

**Wortfrequenzvergleiche
zwischen dem Čechischen und Finnischen**

INAUGURAL - DISSERTATION

zur
Erlangung der Doktorwürde
der Philosophischen Fakultäten
der Albert-Ludwigs-Universität
zu Freiburg i. Br.

vorgelegt
von

Anna-Liisa Seegebrecht
aus Sippola/Finnland

1980

Freiburg 1981

**Wortfrequenzvergleiche
zwischen dem Čechischen und Finnischen**

INAUGURAL - DISSERTATION

zur

Erlangung der Doktorwürde
der Philosophischen Fakultäten
der Albert-Ludwigs-Universität
zu Freiburg i. Br.

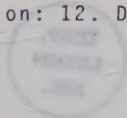
vorgelegt

von

Anna-Liisa Seegebrecht
aus Sippola/Finnland

Freiburg 1981

Referent: Prof. Dr. A. Měšťan
Korreferent: Prof. Dr. E. Weiher
Sprecher: Prof. Dr. F. Buggle
Tag der Promotion: 12. Dezember 1980





Für Barbara und Michael

INHALTSVERZEICHNIS

	<u>Seite</u>
EINLEITUNG	1
I. Abgrenzung der Begriffe der quantitativen Linguistik und terminologische Problematik	
1. Gegenstand und Ziel der quantitativen Linguistik	2
2. Einsatzgebiete der quantitativen Linguistik	4
3. Voraussetzungen für eine quantitative Analyse	6
3.1. Die gezählte Einheit	6
3.2. Auswahl des Materials qualitativ gesehen	10
3.2.1. Das linguistische Kriterium	11
3.2.2. Das psychologische Kriterium	11
3.2.3. Das soziologische Kriterium	13
3.2.4. Das thematische Kriterium	14
3.2.5. Das semiotische Kriterium	19
3.3. Auswahl des Materials quantitativ gesehen	22
3.3.1. Umfang (Korpus)	22
II. Zur Geschichte des Zählens sprachlicher Elemente	
1. Anfänge des Zählens	26
2. Lexikostatistische Arbeiten im Überblick	30
2.1. Frequenzwörterbücher aus dem Bereich der germanischen Sprachen	31
2.1.1. Deutsch	31
2.1.2. Schwedisch	35
2.1.3. Englisch	36
2.2. Frequenzwörterbücher aus dem Bereich der romanischen Sprachen	38
2.2.1. Französisch	38
2.2.2. Spanisch	39
2.2.3. Italienisch	39
2.2.4. Rumänisch	40
2.3. Wortfrequenzuntersuchungen aus dem slavischen Sprachbereich	40
2.3.1. Čechisch	40
2.3.2. Slowakisch	42
2.3.3. Russisch	44

	<u>Seite</u>
2.3.4. Polnisch	45
2.3.5. Altkirchenslavisch	45
2.4. Wortfrequenzuntersuchungen aus dem finno-ugrischen Sprachbereich	47
2.4.1. Finnisch	47
2.4.2. Estnisch	48
2.4.3. Ungarisch	48
III. Lexikostatistische Vergleiche in der tschechischen und finnischen Sprache	
1. Distribution der lexikalischen Einheiten	49
1.1. Klassifizierung nach P. Guiraud	49
1.2. Wortschatzgliederung nach J. Mistrík	50
1.3. Vertikale Wortdistribution des tschechischen und Finnischen	53
1.3.1. Wortfrequenzlisten	53
1.3.2. Vertikale Wortdistribution tschechisch/slovakisch und Finnisch/Estnisch	62
1.3.3. Formale Differenzen	67
IV. Grammatisch-statistische Vergleiche in der tschechischen und finnischen Sprache	
1. Wortklassendistribution	69
2. Stil Kategorien und Wortklassendistribution	76
2.1. Textuale und lexikale Wortklassenverteilung	86
2.2. Stil Kategorien und vertikale Wortdistribution	99
V. Textdeckung	
1. Textdeckung der häufigsten Wörter	102
2. Textdeckung im laufenden Text	112
VI. Zur Semantik der Wörter in der Frequenzliste	
1. Frequenzunterschiede und Polysemie	116
2. Adjektive und ihre Frequenz	119
2.1. Adjektive mit Antonymen	119
2.2. Farbadjektive	120

	<u>Seite</u>
3. Himmelsrichtungen, Wochentage, Monatsbezeichnungen und ihre Frequenz	121
4. Staatsbezeichnungen und Städte- namen und ihre Frequenz	123
4.1. Staatsbezeichnungen mit der Frequenzangabe	123
4.2. Städtenamen und ihre Frequenz	123
5. Numeralia und ihre Frequenz	125
SCHLUBBEMERKUNG	126
ANMERKUNGEN	127
LITERATURVERZEICHNIS	133

EINLEITUNG

Das Hauptanliegen dieser Arbeit besteht darin, das Finnische als nicht-indogermanische Sprache in die vergleichende lexikostatistische und grammatisch-statistische Analyse einzubeziehen und zwar in der konkreten Gegenüberstellung zum Čechischen als indogermanische Sprache. Wahlweise wurden dabei auch das jeweils artverwandte Estnisch bzw. Slovakisch herangezogen. Zunächst wird der Versuch unternommen, eine Übersicht über die Theorienbildung und die Methodik innerhalb der quantitativen Linguistik zu geben.

Schließlich wurde aus dem umfangreichen Forschungsmaterial eine Geschichte des Zählens sprachlicher Elemente erstellt, die auf einer chronologischen Auswahl der für die jeweils fachrelevanten Sprachen repräsentativen Frequenzuntersuchungen basiert.

Die vorliegende Untersuchung erhebt keinen Anspruch auf Vollständigkeit, vielmehr sollten im Hauptteil Unterschiede und Übereinstimmungen in der Semantik der häufigsten Wörter, die Hierarchie der Wortklassen sowie die Textdeckung im Čechischen und Finnischen aufgezeigt werden, um zu weiteren Diskussionen auf dem Gebiet der vergleichenden grammatischen und lexikalischen Statistik anzuregen.

I. ABGRENZUNG DES BEGRIFFES DER
QUANTITATIVEN LINGUISTIK UND
TERMINOLOGISCHE PROBLEMATIK

1. GEGENSTAND UND ZIEL DER QUANTITATIVEN
LINGUISTIK

Um Mißverständnissen vorzubeugen und Ungenauigkeiten zu vermeiden, ist es erforderlich, etwas Klarheit in die terminologische Vielfalt der Fachliteratur zu bringen.

Der britische Mathematiker G.Herdan beispielsweise nannte seine Untersuchung aus dem Jahre 1964, in der er sich mit der Anwendung der statistischen Methoden in den sprachlichen Erscheinungen beschäftigt, "Quantitative Linguistics":

"... Quantitative linguistics appears to be what de Saussure meant by term language as distinct from both langue and parole."¹⁾

Dabei verwendet er auch den Terminus statistische Linguistik, wie ihn später etwa Muller²⁾ in seiner "Sprachstatistik" benutzt oder Cohen³⁾ Anfang der 50er Jahre in seinen Arbeiten "Sur la statistique". Spang-Hansen⁴⁾ schlug auf dem 9. Internationalen Linguisten-Kongreß in Cambridge/Mass. 1962 vor, in Abgrenzung zur non-quantitative linguistics von arithmetical and statistical linguistics zu sprechen. Die sovjetische Forschung hingegen bezeichnet dasselbe als linguistische Statistik.

Těšitelová formuliert den Gegenstand der quantitativen Linguistik so:

"Předmětem kvantitativní lingvistiky je zjišťovat statistické zákonitosti jazykových jevů a jejich vztahů a získávat statistické údaje o jednot-

livých jevech a na jejích základě
charakterizovat analyzované celky".⁵⁾

und als deren Ziel betrachtet sie:

"Kvantitativní údaje, matematické formule,
třídění apod. nejsou cílem kvantitativní
analýzy, alespoň ne pro lingvisty, nýbrž
prostředkem, popr. kontrolou našeho
poznání".⁶⁾

Die quantitative Linguistik ist ein Unterbereich der
mathematischen Linguistik, die sich der quantitativen
Methoden der Statistik, der mathematischen Statistik,
der Wahrscheinlichkeitsrechnung und der Informations-
theorie bedient.

Sie hat sich in verschiedene Objektbereiche differen-
ziert wie etwa die Lexikostatistik oder Phonostati-
stik.

Den zweiten Unterbereich der mathematischen Linguistik
stellt die algebraische Linguistik dar, die mit nicht-
quantitativen Methoden der Algebra, Mengentheorie,
Zeichentheorie, mathematischer Logik, und Topologie
operiert.

Dietrich, Klein⁷⁾ und Těšitelová warnen davor, die quan-
titative und algebraische Linguistik als vergleich-
bare Bereiche der mathematischen Linguistik zu betrach-
ten, da beide verschiedenartig sind und unterschied-
liche Problemstellungen haben. Die algebraische Lingui-
stik befaßt sich mit der Entwicklung formaler Systeme
- zum Beispiel mit der Theorie der mathematisch-lo-
gischen Modelle, die auf der Zeichentheorie beruhen -
zur Beschreibung natürlicher Sprachen als eigener
Theorienbereich innerhalb der Sprachwissenschaft.

Zur Analyse sprachlicher Strukturen hat sich die Statistik als Instrument der quantitativen Linguistik bewährt. Sie hat sich inzwischen in vielen linguistischen Teildisziplinen methodisch etabliert.

2. EINSATZGEBIETE DER QUANTITATIVEN LINGUISTIK

Stenographen, Biologen, Pädagogen und Psychologen hatten längst die quantitativen Methoden auf die sprachlichen Erscheinungen angewandt, bevor die Linguisten sich ihrer bedienten.

Schließlich wurde der VI. Internationale Linguisten-Kongreß in Paris im Jahre 1948 zum Weltforum der quantitativen Linguistik, als unter Vorsitz von B.Trnka ein "Committee on Quantitative Linguistics" gegründet wurde, das speziell für die Probleme dieses jungen Zweiges der Sprachwissenschaft zuständig war.

Die hauptsächlichen Einsatzgebiete der quantitativen Methoden in der Linguistik sind die Lexikostatistik und die grammatische Statistik. Im Bereich der Phonetik hat die Anwendung statistischer Methoden ihre Tradition, ebenso in der Glottochronologie, in der stilistischen Analyse, in der Typologie, in der Stenographie sowie in der Fremd- und Muttersprachendidaktik.

Die Frequenzen der Grapheme und Phoneme spielen heute in der technischen Praxis eine wichtige Rolle - zum Beispiel in der Telekommunikation, in der Druckereibranche oder überhaupt bei der Zusammenstellung von Programmen für die statistische Bearbeitung verschiedenster Sachbereiche, ebenso etwa bei der Dechiffrierung strittiger Autorenschaften der modernen Kriminalistik oder beim Studium der antiken und älteren Literaturen.

Bei komplizierten Analysen wird dabei mehr und mehr als mechanisches Hilfsmittel und als Kontrolleinrichtung der Computer eingesetzt. Er hat aber keinen direkten Einfluß auf die linguistische Theoriebildung.

Die Computerlinguistik ist praktisch so alt wie der Computer selbst. Sie ist ein Anwendungsbereich der Elektronischen Datenverarbeitung (EDV). Der deutsche Terminus ist eine Lehnübersetzung des englischen Computational linguistics oder Computation in Linguistics. Daneben steht häufig der Begriff Linguistische Datenverarbeitung. In den einschlägigen sovjetischen Veröffentlichungen wird auch der Terminus 'Ingenieurlinguistik' (inženernaja lingvistika) oder 'vyčislitel'naja lingvistika', im Estnischen der Begriff 'Infolinguiistik' gebraucht.

Die Computerlinguistik hat in der maschinellen syntaktischen Analyse, in der maschinellen Übersetzung und in der Computerlexikographie eine Dienstleistungsfunktion, über die sie in der Computerlyrik und -graphik hinauswächst.⁸⁾

In den fünfziger Jahren hatte die maschinelle Übersetzung sowohl im englischen als auch im russischen Sprachraum optimistische Befürworter. Chomsky und Bar-Hillel allerdings standen ihr von Anfang an warrend gegenüber. Später wechselte auch Josselson zu den Skeptikern.⁹⁾ Denn eine Wort-für-Wort-Übersetzung läßt sich zwar auf lexikalischer Grundlage durchführen, das Problem der Polysemie jedoch oder des "enzyklopädischen" Wissens, über das ein Übersetzer verfügt, kann mit dem Computer nicht bewältigt werden.

3. VORAUSSETZUNGEN FÜR EINE QUANTITATIVE ANALYSE

3.1. Die gezählte Einheit

Das numerische Grundelement in der lexikostatistischen Analyse ist die gezählte Einheit oder - undifferenziert bezeichnet - das Wort. Bei der Abgrenzung des Begriffs 'Wort' können nach Muller¹⁰⁾ drei Möglichkeiten auftreten, wobei eine graphische Einheit ein Wort bezeichnet, oder diese aus mehreren Wörtern besteht, oder aber mehrere graphische Einheiten ein Wort bezeichnen.

Im allgemeinen definiert man die graphische Einheit als die zwischen zwei Leerräumen oder zwischen einem Leerraum und einem Interpunktionszeichen stehende Zeichenfolge. Sie besteht entweder aus nur alphabetischen oder nur numerischen Zeichen, oder aus Kombinationen dieser Zeichen und/oder Interpunktionszeichen, oder aus Zeichen, die weder dem Alphabet noch Zahlensystem angehören.

In vielen lexikostatistischen Arbeiten der ersten Etappe, so etwa dem Kädingschen Frequenzwörterbuch,¹¹⁾ aber auch in einigen Arbeiten neueren Datums, wie dem Finnischen Frequenzwörterbuch,¹²⁾ sind nur die graphischen Einheiten gezählt worden, die mit alphabetischen Zeichen ausgedrückt werden, während Allén¹³⁾ und Rosengrén¹⁴⁾ alle Zeichen im Text zählten.

Allén teilte die verschiedenen graphischen Einheiten in vier "Worttypen" ein und bezeichnet sie als

1. alphabetisches Wort

2. hybrides Wort (z.B. vgl. und 14,4 etc.)

3. numerisches Wort (z.B. 50 und 1000 etc.)
4. Logogramm (z.B. & und % und § etc.)

Jede Grunddefinition, nach der ein Wort als graphische Einheit fungiert, wird bereits durch die Tatsache in Frage gestellt, daß im Falle der zusammengesetzten Zeiten etwa, wie beispielsweise dem Präteritum im Čechischen 'psala jsem' etc. oder dem Futur im Deutschen 'gehen werden'/'gegangen sein werden', zwei bzw. drei graphische Einheiten auch als 'ein Wort' bezeichnet werden.

Das Wort in der Lexikostatistik ist im Grunde genommen identisch mit der Wortform.

Komplexer stellt sich noch diese Problematik in der Morphologie, wenn es beispielsweise darum geht, ob der Komparativ und Superlativ der Adjektiva zur Positivform gezählt werden soll oder nicht, oder ob Deminutiva und Augmentativa eine eigenständige gezählte Einheit darstellen. Auch erhebt sich die Frage, ob man den bestimmten oder unbestimmten Artikel in der Grund- oder in der modifizierten Form zählt.

Michêa¹⁵⁾ vertritt in diesem Zusammenhang die These, daß Präpositionen, Konjunktionen, Hilfsverben, Partikeln etc. überhaupt nicht gezählt werden sollen.

Herdan führte die Termini der Semiotik 'type' und 'token' in die Linguistik ein, wobei 'type' ein Lexem, eine lexikalische Einheit bezeichnet und 'token' das wiederholte Vorkommen einer Wortform. Těšitelová¹⁶⁾ operiert in ihren lexikostatistischen Arbeiten mit folgenden Termini: 'slovo' (auch 'tvar slova', im Russischen entspr. Sprachgebrauch

'slovoforma') für eine beliebige Einheit eines Textes. 'různý tvar slova' ('různé tvary slova'pl.) für das Vorkommen einer Wortform (z.B. 'Buch', 'Buches', 'kniha', 'knihou' etc.) und 'různě slovo' (lexikální jednotka, lexém) für verschiedene Wörter.

Von Kollektivbezeichnungen ausgehend definiert Muller die von ihm benutzten Termini folgendermaßen:

"Lexik ist ein Terminus der Sprache (langue), (...) den Terminus Lexem wendet man auf die Lexik bildenden Einheiten an, während Vokabular ein Terminus der sprachlichen Äußerung (discour) ist. Werden diese virtuellen Einheiten in der sprachlichen Äußerung aktualisiert, nennt man jede einzelne von ihnen Vokabel und der Terminus Wort bezeichnet jede Okkurrenz einer beliebigen Vokabel (...). Außerdem ist das Vokabular die Gesamtmenge der Vokabeln, die in dem betreffenden Text beliebige Male vertreten sind".¹⁷⁾

Wörter lassen sich - laut Erk¹⁸⁾ - nicht wie Münzen zählen, weil das "monosome" Wort, bei dem ein Eins-zu-eins-Verhältnis zwischen Ausdruck und Inhalt besteht, die Ausnahme und nicht die Regel ist.

Das Wort ist eine Texteinheit, die Vokabel eine lexikalische Einheit. Die Länge des Textes ist in Wörtern angegeben. Jede gesprochene oder geschriebene sprachliche Äußerung kann in Wörter segmentiert werden. Dabei können mehrere Wörter identischer oder veränderlicher Form ein und dieselbe Vokabel repräsentieren. So sind zum Beispiel unter der Vokabel 'Haus' die Belege 'Haus', 'Häuses', 'Häuser' etc. gezählt. Das

wiederum bedeutet, daß Frequenzen in der Regel Eigenschaften von Vokabeln sind. Didaktisch motivierte Frequenzuntersuchungen gehen von einer Gesamtheit der Vokabeln, dem Vokabular, aus.

Die Abgrenzung der Vokabel, die man mit Lemmatisierung bezeichnet, besteht darin, daß man die heterogenen Formen ein und derselben Vokabel zusammenfaßt und die zu verschiedenen Vokabeln gehörenden homographen Formen trennt.

Innerhalb der einzelnen Sprachen bildet sich dabei eine spezielle Problematik heraus, die charatertypologisch oder kontextspezifisch bedingt ist, besonders in Hinsicht auf die Semantik, aber auch auf die Grammatik.

Das Problem der Homonymie läßt sich bei einer Sprache kaum umgehen. Es muß bei jeder lexikostatistischen Analyse berücksichtigt werden, da es die Interpretation ihrer Resultate manchmal empfindlich beeinflußt. Zwei deutsche Ausdrücke sollen die lexikalische Homonymie illustrieren:¹⁹⁾

- | | |
|----------|---|
| 'Bank' | 1. Brettersitz für mehrere Personen nebeneinander |
| | 2. Anstalt des Geldverkehrs |
| 'Strauß' | 1. zusammengebundene Blumen |
| | 2. Kampf |
| | 3. riesiger flugunfähiger Vogel |

Die Homonymie manifestiert sich in den verschiedenen Sprachen auf verschiedenen Gebieten: So tritt in germanischen und romanischen Sprachen verstärkt lexikalische Homonymie auf, während in den slavischen Sprachen die morphologische Homonymie dominiert.²⁰⁾

3.2. Auswahl des Materials qualitativ gesehen

Für die Auswahl des Materials gibt es keine genaue Regel, und von Polenz postulierte im Jahre 1967, daß der Sprachforscher

"... eine wohlüberlegte Auswahl zu treffen" habe, "wenn er der Gefahr entgehen will, nur den zufälligen Funden aus bevorzugter Lektüre oder den einseitigen Perspektiven eines Sprachbewußtseins nachzugehen".²¹⁾

Dokumentation einerseits und grammatische, lexikologische, syntaktische Arbeit andererseits müssen sich nach Hellmann²²⁾ also von Stufe zu Stufe quasi gegenseitig "hinaufschaukeln".

Die qualitative Auswahl des Materials unterliegt auch dem Ziel der Arbeit, was bei der statistischen Analyse verfolgt wird. Daher unterscheidet sich ein Frequenzwörterbuch, das für den schulischen Gebrauch bestimmt ist, von einem wissenschaftlichen.

Unter diesem Gesichtspunkt stellte das Mannheimer Institut für deutsche Sprache folgende Kriterien auf:

1. das linguistische Kriterium
2. das psychologische Kriterium
3. das soziologische Kriterium
4. das thematische Kriterium
5. das semiotische Kriterium

3.2.1. Das linguistische Kriterium

Hierbei handelt es sich um eine von der Funktion abhängige Konstellation, um Charakteristika

- a) einer Sprache überhaupt, z.B. des Čechischen, des Deutschen, des Finnischen etc.
- b) der einzelnen Funktionsstile einer Sprache, des künstlerischen, journalistischen, administrativen, wissenschaftlichen etc. Stils
- c) der Sprache einzelner Autoren, d.h. ihres Wortschatzes, beziehungsweise ihres Vokabulars, zum Beispiel Puškins,²³⁾ des bulgarischen Lyrikers Botev²⁴⁾ oder Trakls²⁵⁾
- d) der Sprache einzelner Textteile²⁶⁾

3.2.2. Das psychologische Kriterium

Von Bedeutung ist das Lebensalter dessen, der eine Sprache erlernt. Für jede Alterskategorie (Kindes-, Schul-, Erwachsenenalter) muß die Sprachdidaktik das lexikalische und grammatische Minimum entsprechend bestimmen. Auch ist es wesentlich, ob es sich um das Erlernen der Muttersprache oder einer Fremdsprache handelt.

Das Frequenzwörterbuch von E.A. Štejnfeldt,²⁷⁾ das als objektive Grundlage für die Festlegung des lexikalischen und grammatischen Minimums der russischen Sprache in estnischen Schulen bestimmt ist, stellt in dieser Hinsicht ein Modell dar. Die Texte -

insgesamt 400 000 Wörter - aus denen dieses Wörterbuch entstanden ist, stammen aus fünf Stilgruppen:

1. schöngeistige Literatur für Kinder im Schulalter (100 000 Wörter = 25 %)
2. schöngeistige Literatur für Erwachsene, Erzählungen aus Zeitschriften (50 000 Wörter = 12,5 %)
3. Theaterstücke (50 000 Wörter = 12,5 %)
4. russische Sendungen für Jugendliche im estnischen Rundfunk (100 000 Wörter = 25 %)
5. Zeitungsartikel aus verschiedenen Zeitschriften und Zeitungen für Jugendliche (100 000 Wörter = 25 %)

In diesem Frequenzwörterbuch geht es vorrangig um die für Jugendliche bestimmte geschriebene und gesprochene Sprache.

Wie das Čechische Frequenzwörterbuch von Jelínek Bečka und Těšitelová²⁸⁾ zeigt, existiert hinsichtlich der Frequenz der Wörter oder Frequenz der grammatischen Kategorien kein grundlegender Unterschied zwischen der schöngeistigen Literatur insgesamt und der Literatur für Jugendliche.

Vergleicht man die Frequenz des in Štejnfeldt's Frequenzwörterbuch gebräuchlichsten Substantivs 'reb-jata' auf Rang 51 mit dem Rang 439 desselben Wortes im Frequenzwörterbuch von Zsorina²⁹⁾, so basiert die Frequenzverschiebung auf der spezifischen Auswahl der Texte. Ähnliches läßt sich an nachfolgenden Beispielen beobachten:

Wort	Štejnfeldt	Zasorina
škola	65	418
mama	158	334
mal'čik	160	752
devočka	161	1077

3.2.3. Das soziologische Kriterium

Bei der qualitativen Auswahl des Materials wurde bisher auch vom soziologischen Kriterium nicht der Stellenwert zugemessen, der ihm aufgrund seiner Wichtigkeit entsprochen hätte.

So wurde beispielsweise bei der Textauswahl für eine lexikostatistische Analyse nur ganz vereinzelt berücksichtigt, ob es sich dabei um die Mitteilungen von Männern oder Frauen handelte.

In Bezug auf die geschlechtsspezifische Relevanz der Wortfrequenz will Těšitelová³⁰⁾ festgestellt haben, daß in monologischen wissenschaftlichen Texten von Männern eine sichere Ausgewogenheit im Wortschatz herrscht, unabhängig von der Länge des Textes. In analogen Texten von Frauen könne man angeblich keine größere Konzentriertheit auf die Thematik nachweisen. Ansonsten sei die Auswahl der Wörter und ihre Distribution natürlich weitgehend eine Sache des Bildungsgrades.

Bei der Untersuchung der Frage, ob ein Zusammenhang zwischen den individuellen Gegebenheiten und der quantitativen Prägung des Stils besteht, stellte Schröder³¹⁾ fest, daß sich anhand signifikanter Korrelationen Hinweise auf eine Bevorzugung adjektivischer Aussageformen bei den weiblichen Versuchspersonen ergaben. Die männlichen Teilnehmer hingegen verwendeten umfangreichere Wörter und bevorzugten Verben, Adverben, Possessivpronomina.

Für die Intelligenz wurden signifikante Beziehungen hauptsächlich zur Frequenz der Wortarten ermittelt, wobei die Verben mit der intellektuellen Leistungsfähigkeit in negativer, die Adverbien und Adjektiva dagegen in positiver Korrelation standen. Bei den Charaktermerkmalen ergaben sich dagegen mehr Zusammenhänge mit dem Umfang der Texte, Sätze und Wörter.

3.2.4. Das thematische Kriterium

Von Polenz fordert eine Berücksichtigung gerade der

"Mittelschicht der sachlichen Informationssprache des öffentlichen Sprachgebrauchs in Politik, Verwaltung, Wirtschaft und Technik".

Man sollte sich nicht

"der Gefahr aussetzen, daß spätere Linguistengenerationen auch (...) die Arbeit bemängeln müssen mit dem Vorwurf, man hätte - trotz moderner sprachwissenschaftlicher Einsichten, Methoden und Mittel - die häufigsten und alltäglichsten Erscheinungen (...) zu selbstverständlich, stilistisch problemlos oder

etwa normwidrig erscheinen lassen (...). Eine repräsentative Quellenauswahl für die Gegenwartssprache sollte also neben der ständig variablen Liste der z.Z. für bedeutend gehaltenen Schriftsteller auch eine der sprachsoziologischen Wirklichkeit angemessene Menge von Texten aus Tageszeitungen erfassen".³²⁾

Das Institut für deutsche Sprache in Mannheim legte bei der qualitativen Auswahl eines Korpus bei seiner Untersuchung auf die Betonung auf das Sachsystem, praktisch etwa auf Sachkataloge in Bibliotheken. Die sachbezogene Klassifizierung erfolgte

1. im Rahmen verschiedener Stilgruppen, d.h. Gattungen
2. nach der Erscheinungsform

Zu 1. Vorgeschlagen wird die Verteilung des Korpus auf vier Gattungen:

1. Belletristik
 - a) Dichtung (Künstlerische Weltgestaltung)
 - b) Gehobene Unterhaltung(sromane)
 - c) Trivialliteratur
2. Informierende (unterrichtende) Literatur
 - a) Wissenschaftliche und populärwissenschaftliche Literatur
 - b) Memoiren
 - c) Aktuelle Information
3. Anleitende (belehrende) Literatur (Gebrauchsliteratur)
 - a) Fachliche Anleitung
 - b) Gebrauchsanweisungen

- c) Kochbücher
 - d) Gesundheits-Hausbücher
 - e) Bastelbücher
 - f) Reisebücher (-führer)
4. Werbende und auffordernde Literatur
- a) Kommerzielle Werbung
 - b) Politische Werbung
 - c) Soziale und weltanschauliche Werbung

Die Verteilung des Korpus auf Erscheinungsformen wird folgendermaßen vorgeschlagen:

- 1. Bücher
 - a) mit einheitlichem Inhalt
 - b) mit verschiedenen Inhalten (Sammelwerke)
- 2. Broschüren
- 3. Periodika
 - a) Zeitungen (Tages- und Wochenzeitungen)
 - b) Zeitschriften
 - c) "Hefte"
- 4. Drucksachen
 - a) kommerzielle
 - b) öffentliche
 - c) private
- 5. Plakate

Die die Thematik berücksichtigende qualitative Auswahl der Texte unterliegt naturgemäß der jeweiligen Zielsetzung, ob es sich beispielsweise um die Erstellung eines Frequenzwörterbuches einer Literatursprache handelt oder um die Erstellung einer Wortfrequenzliste einer Fachsprache.

Theoretisch müßte man, um die Frequenz der Wörter einer Sprache zu ermitteln, alle Texte dieser Sprache, geschriebene wie gesprochene, analysieren, und dies auch im Hinblick auf ihre Thematik. Daher ist eine Auswahl unerläßlich.

Allgemein lassen sich hier zwei Methoden unterscheiden:

1. Es wird eine bestimmte Anzahl der kompletten Texte aus verschiedenen Themenbereichen ausgewählt und exzerpiert und die in ihnen vorkommenden Wörter nach ihrer Frequenz sortiert, oder aber
2. werden mehrere kürzere Textabschnitte aus einer möglichst großen Anzahl thematisch-verschiedener Quellen ausgesucht und nur die Textteile exzerpiert und die in ihnen vorkommenden Wörter nach ihrer Frequenz bestimmt.

Die erste Art der Auswahl bezeichnet Těšitelová: výběr "co do hloubky" und die zweite: výběr "co do šířky".³³⁾

Bei den relativ hochfrequentierten Wörtern hängt das Ergebnis nicht so empfindlich von der Größe der Teilmenge ab, wie bei den niedriger frequentierten. Es kann ein Mißverhältnis entstehen, wenn die Auswahl nicht umfangreich genug ist.

Billmeier³⁴⁾ unternahm den Versuch, das Verhältnis eines Teiltextes zum Gesamttext auszudrücken und diese Beziehung als Maß für die Genauigkeit des Teiltextes - bezogen auf den Gesamttext - oder als Grad der Abhängigkeit des Teiltextes zu definieren.

Er zeigt die Verkleinerungsgröße eines Teiltextes auf, die die Wortverteilung, die Grammatik und die Syntax des Gesamttextes widerspiegelt und damit verbindliche Aussagen über den Korpus zuläßt, und kommt dabei zu dem Ergebnis, daß eine Vervielfachung des Materials nicht im gleichen Maße eine Vervielfachung der Genauigkeit mit sich bringt. Außerdem hängt der gewählte Genauigkeitsgrad vom Untersuchungsziel ab. Je seltener eine sprachliche Erscheinung im Gesamttext auftritt, desto größer muß der Wert der Wahrscheinlichkeit gewählt werden.

Auswahltexte um oder unter 10 000 Wörter können nicht als verbindlich für den Gesamttext angesehen werden. Korpora dieser Größe müssen komplett exzerpiert werden.

Hellmann betont in diesem Zusammenhang, daß bei einer Auswahl - statistisch gesehen - grundsätzlich 10 Prozent aus 100 Titeln immer besser seien als 100 Prozent aus 10 Titeln. Die Zahl der Titel kann also verhältnismäßig hoch gewählt werden.

Für beide Methoden gilt der Grundsatz einer möglichst mannigfaltigen Verteilung des Korpus auf Sachgebiete.

Hellmann schlägt hier zehn Sachgebiete vor:

1. nicht sachgebietsgebundene (Fiction)
2. Politik
3. Wirtschaft (Industrie, Landwirtschaft etc.)
4. Technik
5. Kultur
6. Kunst
7. Soziales

8. Naturwissenschaften (Physik, Chemie, Biologie, Medizin)
9. Geisteswissenschaften (Rechts- und Literaturwissenschaften, Theologie, Philosophie, Geschichte)
10. Sport

Vergleicht man nun die Sachgebietsauswahl des Čechischen, slowakischen, russischen und deutschen Frequenzbuches (= FWB), so zeigt sich, daß das Čechische Frequenzwörterbuch das mannigfaltigste Themenspektrum besitzt:

FWB	1	2	3	4	5	6	7	8	9	10
čechisch	x	x	x	x	x		x	x	x	
slovakisch	x	x			x		x	x	x	
russisch	x	x			x		x	x	x	
deutsch	x	x	x		x			x	x	

3.2.5. Das semiotische Kriterium

Nach dem Schema der Grundfunktionen des Sprachzeichens von Bühler³⁵⁾ existieren drei aus dem Wesen der Sprache selbst sich ergebende Funktionen:

1) die darstellende, 2) die expressive und 3) die appellative Funktion der Sprache. Diese reichen für jede mitteilende Äußerung aus. Mukařovský jedoch fügt eine vierte Funktion zu dem Bühler'schen Schema hinzu, und bezeichnet sie als ästhetische Funktion der Sprache. Die ästhetische Funktion

"stellt nämlich in den Mittelpunkt des Interesses die Komposition des Sprachzeichens, während die erstgenannten drei zu außersprachlichen Instanzen und zu Zielen tendieren, die das Sprachzeichen überschreiten".³⁶⁾

Auf der Grundlage dieser Vierteilung der Sprache nach Mukařovský nimmt Kern eine qualitative Auswahl der Texte nach semiotischen Kriterien vor.

Das semiotische System übernimmt er von W. Morris,³⁷⁾ ohne von dessen Einteilung der sprachlichen Typen Gebrauch zu machen.

Seiner Meinung nach kann man die ganze Textvielfalt nicht in ein System einspannen, wie dies in Bibliotheken und Bücherkatalogen der Fall ist. Die bisher vorggeführten Klassifikationsbemühungen krankten alle daran, daß sie ihre Kriterien ausschließlich aus den Texten selbst ableiteten und damit Ergebnisse vorwegnahmen.

Kern³⁸⁾ versucht, "eine Typik der geschriebenen Sprache entsprechend der Schreibintention" in Anlehnung an das bekannte Kommunikationsdreieck (es-du-ich) zu entwerfen: orientiert erstens am Objekt, zweitens am Kommunikationspartner, drittens an sich selbst.

Demgemäß klassifiziert er die Texte in:

A) sachorientierte

- 1) Analyse
- 2) Beurteilung
- 3) Anwendbarkeitsprüfung
- 4) Systematisierung
- 5) ungezieltes Experiment

- B) du-orientierte
 - 1) Information
 - 2) Überredung
 - 3) Appell
 - 4) Belehrung
 - 5) Unterhaltung
- C) ich-orientierte
 - 1) Bestandsaufnahme
 - 2) Bewertung
 - 3) Auswertung
 - 4) Reflexion
 - 5) "Gedanken"spielerei

wobei es zu Überschneidungen kommt:

- z.B. A3-B1 (Anwendbarkeitsprüfung - Information)
ergibt die Konkretisierung Gebrauchsanweisungen und Rezepte, oder
- A2-B4 (Beurteilung - Belehrung)
Lehrbücher
- C4-B1 (Reflexion - Information)
autobiographische Bemerkungen, die meisten Privatbriefe.

In der Kombination der einzelnen Komponenten dominieren immer einige von ihnen und bilden eine subjektive Modifikation der zugehörigen Texte . - Es handelt sich dabei vor allem um einen theoretischen Entwurf.

Kučera und Francis³⁹⁾ wählten die Texte für ein Frequenzwörterbuch des heutigen amerikanischen Englischen nach Funktionsstilen und semiotischen Aspekten aus. Sie unterscheiden sie dabei z.B. innerhalb des Journalismus, die Reportage, den Leitartikel, die Rezension.

Ein spezielles Problem bei der qualitativen Auswahl der Texte stellt das Einordnen von Dialekt- oder Slangwörtern dar. Diese werden im allgemeinen aus dem Korpus aussortiert, es sei denn, es handelt sich um ein spezielles Dialekt- bzw. Slangwörterbuch.

Strittig ist auch die Frage, ob man Lehrbücher mit ihrer spezifischen Sprachkonstellation in das Korpus impliziert oder nicht, wie dies im speziellen Fall von Schulfrequenzwörterbüchern seine Berechtigung hat.

So stellte Wood⁴⁰⁾ beispielsweise aus dem Wortschatz von 16 Französisch-Lehrbüchern seine Wortfrequenzliste zusammen.

Für die Auswahl des Materials nach qualitativen Gesichtspunkten gibt es also keine bestimmten Richtlinien und Regeln. Sie unterliegt stets der subjektiven Meinung des Forschers.

Daher ist die Zusammenarbeit der Disziplinen für eine optimale Problemlösung unerlässlich.

3.3. Auswahl des Materials quantitativ gesehen

3.3.1. Umfang_(Korpus)

Zur Umfangbestimmung der Korpora sind bislang keine allgemeingültigen "Maße" definiert worden. Der Umfang der Materialauswahl, auch Stichprobenumfang genannt, ist die Anzahl jener Einheiten, die im konkreten Fall analysiert werden.

Alekseev⁴¹⁾ erscheinen als Mindestumfang eines Textes für statistische Auszählungen 1 000 Wörter ausrei-

chend zu sein, aber um Gesetzmäßigkeiten feststellen zu können, empfiehlt er größere Wortmengen von 50 000, 150 000, 200 000 Wörtern heranzuziehen.

Hoffmann wiederum hält 35 000 Wörter für günstig und effektiv, nachdem ihm Untersuchungsergebnisse für etwa 15 Fachsprachen vorlagen.⁴²⁾

Frumkina,⁴³⁾ Epstein⁴⁴⁾ und Guiraud⁴⁵⁾ fordern zur Gewährleistung der Reliabilität der ersten 1 000 häufigsten lexikalischen Einheiten 400 000 bis 1 000 000 Wörter.

Der Maximalumfang des Materials wird von Přihoda⁴⁶⁾ auf 2,5 Millionen Wörter festgesetzt, während Těšitelová ca. 1,5 Millionen Wörter als Maximum bezeichnet, wie sie sie mit Bečka und Jelinek für das tschechische Frequenzwörterbuch bearbeitete.

Frumkina stellte fest, daß die Wortfrequenz der verschiedenen Frequenzwörterbücher mit unterschiedlicher Genauigkeit angegeben worden ist. Um ein möglichst präzises Ergebnis zu erzielen, bestimmte sie den sogenannten "relativen Fehler" der Frequenz.

Anhand des Materials des spanischen Frequenzwörterbuches von Hoz und des russischen Frequenzwörterbuches von Josselson entwickelte sie folgende Formel:⁴⁷⁾

$$\delta = \frac{z_p}{N_p}$$

Dabei bedeutet N = Stichprobenumfang

δ = gegebener relativer Fehler

p = Frequenz, die in der gegebenen Stichprobe mit dem Fehler festgesetzt sein soll und die den gegebenen relativen Fehler nicht übersteigt.

Z_p = konstant

Ferner postulierte sie:

- 1) das Frequenzwörterbuch muß die Wortfrequenz mit genügender Genauigkeit anführen,
- 2) die Auswahl darf nicht einen bestimmten Umfang überschreiten, damit man sie noch bearbeiten kann
- 3) es ist zulässig, daß einige Wörter, die im Text vorkommen, nicht in das Frequenzwörterbuch gelangen. Ihr Prozentsatz muß allerdings unter 0,5 % liegen.

In den Frequenzwörterbüchern zugrundegelegte
Wortmassen:

Sprache und Autor des Frequenzwörter- buches	Anzahl der Wörter
Russisch: Josselson	1 000 000
Šteijnfel'dt	400 000
Zasorina	1 056 382
Deutsch: Käding	10 910 777
Rosengrén	2 976 894
Čechisch: Jelínek, Bečka Těšitelová	1 623 527
Slovakisch: Mistrík	1 000 000
Schwedisch: Allén	1 000 000
Finnisch: Saukkonen, Haipus, Niemiörpi, Sulkala	408 301

II. ZUR GESCHICHTE DES ZÄHLENS

SPRACHLICHER ELEMENTE

1. DIE ANFÄNGE DES ZÄHLENS

Das Zählen der sprachlichen Einheiten diente ursprünglich Zwecken, die in erster Linie nichts mit linguistischen Problemen zu tun hatten, eher mit ökonomischen, politischen, soziologischen o.ä.

In Zeiten, als geheime Botschaften noch auf das geschorene Haupt eines zuverlässigen Sklaven geschrieben und derselbe nach dem Nachwachsen der Haare zum Bestimmungsort gesandt wurde, als für Geheimschriften, wie etwa für die des orientalischen Altertums, die in Jerämia angewandt worden ist, das Alphabet umgedreht (d.h. der letztere Buchstabe den ersten, der vorletzte den zweiten usw. vertrat), oder als bei altjüdischen Amuletten jeder Buchstabe des Alphabets durch den folgenden ausgedrückt wurde, hat man für die Entschlüsselung der Geheimschrift (Kryptographie) keine Frequenzuntersuchungen der sprachlichen Einheiten gebraucht.¹⁾

Aber im Jahre 1474 scheint Sicco Simonetta, der Ratgeber dreier Herzöge von Sforza²⁾ und ein einschlägiger Kenner des damaligen Chiffrenwesens, Latein vom Italienischen durch die Buchstabenfrequenzen unterschieden zu haben, obwohl er nicht ausdrücklich die Frequenzen bestimmt hatte.

Obri gens machte man sich im Frankreich Karls IX, an dessen Hof der berühmte Kryptologe Vigenère wirkte, einen Spaß daraus - etwa dem heutigen Kreuzworträtsel vergleichbar -, sich mit Kryptographie zu beschäftigen.

Seit dem vorigen Jahrhundert hat man die Statistik benutzt, um Geheimschriften zu dechiffrieren. Man

zählte dabei nicht nur die Frequenz eines einzelnen Buchstaben, sondern auch die Frequenz zweier (Bigramm) und dreier (Trigramm) Buchstaben.³⁾

Im 19. Jahrhundert wurden Spracheinheiten statistisch erfaßt, um zweckmäßige Schreib- und Typographenmaschinen zu entwickeln, oder um ein effektives stenographisches System auszuarbeiten. Das letztere hatte Kädig⁴⁾ im Auge, als er für seine bahnbrechende Untersuchung Wörter zählte.

Bacmeister⁵⁾ wiederum rechnete dem kaiserlichen Hof vor, wieviel Geld man beim Telegraphieren nach Amerika sparen würde, wenn ein von ihm ausgearbeitetes Draht-Alphabet auf der Grundlage der Buchstabenfrequenz eingeführt werden würde.

Der irische Buchdrucker Gamble⁶⁾, der als Missionar der Presbyterianischen Kirche in China wirkte, berechnete die relative Frequenz der chinesischen Wörter, um eine möglichst effektive Druckmaschine zu entwickeln, damit für Bibelübersetzungen nur die meistgebrauchten chinesischen Schriftzeichen benutzt würden.

Auch Morse und Braille mußten, um ihre Verständigungssysteme aufbauen zu können, von der Frequenz-zählung der sprachlichen Einheiten ausgehen - Morse für sein aus Strichen und Punkten bestehendes Alphabet und Braille für die Blindenschrift.

Später zählte Knowles⁷⁾ 100 000 englische Wörter nach ihrer Frequenz, um ein möglichst einfaches, zweckmäßiges System für die englische Blindenschrift auszuarbeiten.

Es kommt nicht von ungefähr, daß die wenigen Untersuchungen zur statistischen Stilistik in naturwissenschaftlichen Zeitschriften erschienen sind und dadurch meistens den Sprachwissenschaftlern unbekannt blieben. Denn die Sprachwissenschaft allgemein bezog seinerzeit - sehr im Gegensatz zu anderen Wissenschaftsdisziplinen - die Statistik nicht als Hilfsmittel zur Untermauerung ihrer Forschungsergebnisse ein.

Die Situation änderte sich auch nicht, als neben die diachron-ausgerichtete Linguistik ein gleichberechtigter synchroner Forschungszweig getreten war. Erst als psychologische und soziologische Fragestellungen in die Sprachanalyse Eingang fanden, trat das Problem der Frequenz bestimmter sprachlicher Erscheinungsformen in den Vordergrund des wissenschaftlichen Interesses.

Förstemann hoffte im Jahre 1852, daß sich

"durch statistische Angaben über das Vorkommen der einzelnen Laute Resultate über die Entwicklung der Sprachen und über das Verhältnis der einzelnen Idiome zueinander"

erzielen lassen und man dadurch

"der Erkenntnis des Sprachgeistes und Sprachlebens näher komme"⁸⁾.

In seiner Untersuchung über die Lautverhältnisse der griechischen und lateinischen und deutschen Sprache zählte er auch den durchschnittlichen Vokal- und Konsonantengebrauch anderer Sprachen aus und kam zu

dem Ergebnis, daß das Čechische durchschnittlich 188 Konsonanten auf 100 Vokale benutzt, während das Finnische 96 Konsonanten, das Deutsche 177 und das Russische 150 verwendet.⁹⁾

Der britische Mathematiker und Logiker De Morgan¹⁰⁾ stellte sich vor, daß ein stilistisches Problem objektiviert werden könnte, indem man die Wortlänge als stilistisches Kriterium eines Autors untersuchte.

Er selbst jedoch verifizierte seine Angaben noch nicht empirisch, und erst Mendenhall¹¹⁾ hob - an De Morgans Vorschläge anknüpfend - die Wortlängenverteilung als Charakteristikum des Individualstils mehrerer Schriftsteller heraus und versuchte, auch Fragen umstrittener Autorenschaft zu klären.

In seinen Untersuchungen "Who was Junius?" (Stockholm 1962) und "A Statistical Method for Determining Authorship, The Junius Letters 1769 - 1772" (Göteborg 1962) ging Ellegård in exemplarischer Weise der Lösung dieses Problems nach. Er analysierte mit Hilfe von quantitativen Methoden Sprache und Stil der Junius'schen Briefe¹²⁾ und kam durch den Vergleich der relativen Frequenz des in diesem vorkommenden Wortschatzes mit dem der journalistischen Literatur der damaligen Zeit zu dem bereits vermuteten Resultat, daß Junius ein Beamter des Kriegsssekretariats namens Sir Philipp Francis war.

Im Institut der čechischen Sprache (Ústav pro jazyk český ČSAV) in Prag wurden, um ein weiteres Beispiel zu nennen, die Handschriften "Rukopis královédvorský" und "Rukopis zelenohorský" mit statistischen Methoden analysiert. Das Autorenteam Těšitelová, Nebeská und

Králík¹³⁾ verglichen den Wortschatz der Handschriften mit dem der Texte der tschechischen Wiedergeburt und der authentischen, alttschechischen Dalimil-Chronik. Dabei ergab sich, daß beide Handschriften mit ihren quantitativen Eigenschaften die gleiche Konstellation reflektierten wie die der Texte der tschechischen Wiedergeburt.

2. LEXIKOSTATISTISCHE ARBEITEN IM OBERBLICK

In der Entwicklung der Lexikostatistik unterscheidet man zwei Etappen. In der ersten - vom Anfang des Jahrhunderts bis in die 60-er Jahre - waren pädagogische, psychologische, stenographische Gesichtspunkte vorrangiger als linguistische. In dieser Zeit entstanden die großen Frequenzwörterbücher des Deutschen, Englischen und Französischen. Von arbeitstechnischer Seite her betrachtet haben sie eines gemeinsam: ihr Wortmaterial ist manuell sortiert.

In der zweiten - unmittelbar in die Gegenwart hineinreichenden - Etappe wurde die Lexikostatistik vor allem zum gemeinsamen Forschungsgegenstand von Linguisten und Mathematikern, die sich die Vorteile der Computer-Technik zunutze machten.

Im folgenden sollen nun jene Frequenzwörterbücher vorgestellt werden, die von entscheidender wissenschaftlicher bzw. didaktischer Relevanz sind.

2.1. Frequenzwörterbücher aus dem Bereich der germanischen Sprachen:

2.1.1. Deutsch

Das erste Frequenzwörterbuch im Sinne der heutigen Frequenzwörterbücher war das im Jahre 1897 unter dem Titel "Häufigkeitswörterbuch der deutschen Sprache" (Berlin-Steglitz 1897) erschienene, von Käding, dessen Umfang erst in den 50-er Jahren durch die maschinelle Auszählung übertorffen werden konnte. Kädings Ziel war es, "ein wirklich zweckmäßiges stenographisches System aufzustellen".¹⁴⁾ Bei der Textauswahl legte er daher das Hauptgewicht auf die fachsprachliche Literatur und entnahm das spezifische Material aus dem juristischen, kaufmännischen, theologischen, militärischen und belletristischen Bereich. Den Materialumfang seiner Untersuchung, in der er 10 910 777 Wörter zählte, begründete er damit, daß die Frequenz eines bestimmten Wortes je nach Stoff den größten Schwankungen unterworfen sei und diese Schwankungen umso geringer wären, je größer der gesamte verarbeitete Zählstoff sei.

Er zählte die Wörter in der Form, in der sie im Kontext vorkommen.

Die Frequenzangaben betreffen nicht nur Wörter, sondern auch kleinere Einheiten: Silben, Wortstämme, Präfixe, Suffixe, Buchstaben und Laute. Im gesamten Material kommen 258 173 verschiedene Wortformen vor, davon 126 862 (49,14 %) nur ein einziges Mal.

Käding hob schon die Tatsache hervor, daß eine Sprache mit einem überraschend geringen Wortschatz verstanden und verwendet wird. Die 15 häufigsten Wörter

decken 25,22 % des gesamten Materials der deutschen Sprache ab, mit den 66 häufigsten Wörtern werden 50,06 % und mit 320 72,25 % abgedeckt.

Theoretisch müßte man mit einem Wortschatz der 320 häufigsten Wörter 75 % jedes beliebigen Textes verstehen; was jedoch nicht heißt, daß 75 % des Inhaltes verstanden werden, da ein großer Teil der Wörter keine konkrete Bedeutung hat und bestimmte ungewöhnliche Wörter in manchen Texten die Schlüsselposition einnehmen.

Das Kädingsche Frequenzwörterbuch wurde zum Ausgangspunkt für weitere Wortfrequenzuntersuchungen:

B.Q. Morgan wählte in seinem "German Frequency Word Book" (New York 1928) die 2 400 meistgebrauchten deutschen Wörter nach den Kädingschen Frequenzlisten aus und verfaßte damit ein Hilfsmittel für Deutsch-als-Fremdsprache.

Vom Kädingschen Frequenzwörterbuch ging auch H. Bakonyj in seinem Werk "Die gebräuchlichsten Wörter der deutschen Sprache" (München 1934) aus. Er hebt einen Korpus von 5 874 Wörtern aus der Kädin'g'schen Totalmasse heraus und bestimmt deren absolute und relative Frequenz sowie die "Wertstufe", d.h. den Wichtigkeitsgrad eines Wortes, denn seiner Meinung nach reichen für didaktische Zwecke Frequenzangaben allein nicht aus.

Auch H. Meier benutzte in seiner "Deutsche Sprachstatistik" (Hildesheim 1964) Kädin'g als Ausgangspunkt. Er unterteilt den deutschen Wortschatz in 12 Frequenzstufen und versucht, ein nach Verfahren der Dezimalklassifikation entworfenes Frequenznetz auf-

zustellen, das den schriftssprachlichen Gebrauchs-
wortschatz in 100 Felder aufgliedert. Ähnlich, der
späteren horizontalen Wortdistribution von Mistrík,¹⁵⁾
teilt Meier die Frequenzliste in Gruppe A und Gruppe
B ein. Die Wörter der letzteren bezeichnete Daniels¹⁶⁾
als "Voraussetzungen allen Denkens und Seins, an de-
nen man (...) die indogermanische Struktur des Deut-
schen in seltener Eindeutigkeit ablesen kann".¹⁷⁾

Mit seinem Frequenzwörterbuch hoffte er, ein Hilfs-
mittel geschaffen zu haben, nicht nur für Sprachthe-
oretiker und -praktiker, Sprach- und Entwicklungs-
psychologen, sondern auch für Pädagogen, Religions-
wissenschaftler, insbesondere Bibelforscher, Litera-
turwissenschaftler und Soziologen.

Die 40-jährige Arbeit von Meier ist durchaus beach-
tenswert, allerdings bleibt dem Benutzer das Werk
recht unzugänglich, denn erstens stellt er weder
die Prinzipien der Umgestaltung des Käding'schen
Frequenzwörterbuches ausführlich dar, noch erläutert
er die Arbeitsverfahren. Auch erfährt man die exak-
te zeitliche Begrenzung der exzerpierten Texte nicht.
Zudem sind die Wortfrequenzlisten teilweise über-
holt, z.B. fehlen schon die für die 60-er Jahre ak-
tuellen Wörter, wie 'Auto', 'Führerschein', 'Beton',
'Flugzeug'.

Schließlich will W.D.Ortmann mit seiner breit ange-
legten Untersuchungsreihe¹⁸⁾ zeigen, daß Kädings
Zählungen "nicht als veraltet, einseitig schrift-
sprachlich, abseits wichtiger Sachbereiche ange-
siedelt und daher didaktisch irrelevant"¹⁹⁾ bezeich-
net werden können.

H.-H. Wängler wählte in seinem "Rangwörterbuch hochdeutscher Umgangssprache" (Marburg 1963) für seine Auszählungen zwei gleichgroße Mengen (je 80 266 Wörter) aus, und zwar sowohl aus Quellen der deutschen geschriebenen als auch der gesprochenen Sprache. Als Quelle für das geschriebene Deutsch dienten ihm Tagespresse und Wochenmagazine, für das gesprochene auf Tonband aufgenommene ungezwungene Alltagsgespräche. Als Ergebnis seiner Arbeit entstanden vier Listen, eine mit Angaben der Wortfrequenz in alphabetischer Reihenfolge, dann eine zweite, die die Frequenz des gesamten Materials widerspiegelt, eine dritte über die Frequenz des geschriebenen Materials und eine vierte über die des gesprochenen Materials.

I. Rosengrén hat ein 2-bändiges "Frequenzwörterbuch der deutschen Zeitungssprache"²⁰⁾ auf der Grundlage der "Welt" und der "Süddeutschen Zeitung" (1.11.1966 - 30.10.1967) erstellt. Das gesamte Material betrug dabei 2 976 894 Wörter aus 6 595 Artikeln, geschrieben von 594 Autoren, und wurde in die fünf Kategorien

1. Meinung
2. Politik
3. Kultur
4. Wirtschaft
5. Vermischtes

unterteilt.

Während im Band 1 das Material nach Wortformen aufgelistet ist, basiert der Band 2 auf einer Lemmatisierung des gesamten Materials des ersten Bandes.

Rosengrén's Meinung nach ist es wichtig, aus Homogenitätsgründen die Nachrichtentexte der Tagespresse

zu beachten, die die Entwicklung der Sprache am schnellsten und nicht unwesentlich beeinflussen.

Von spezieller Art sind die Wortfrequenzlisten von L.Hoffmann,²¹⁾ die - vornehmlich für ausländische Studenten und Fremdsprachenlehrer bestimmt - den Fachwortschatz (jeweils 1 100 Wörter) der Physik, der Chemie und Pädagogik (russische Terminologie) zum Gegenstand haben. Analoge Untersuchungen gibt es in der Sowjetunion, wie etwa die von R.G. Piotrovskij herausgegebene "Statistika teksta" I (Minsk 1969) und von A.I. Kiselovskij herausgegebene "Statistika teksta" II (Minsk 1970).

H. Erk stellte - ebenfalls für den Unterricht in Deutsch-als-Fremdsprache - im Rahmen seiner Untersuchungen zur "Lexik wissenschaftlicher Fachtexte" Frequenzlisten nach den Wortarten Substantiv, Verb, Adverb zusammen und legte ein Gesamtkorpus von 250 000 Wörtern zugrunde.²²⁾

2.1.2. Schwedisch

S. Allén wählte in seinem "Nusvensk frekvensordbok baserad på tidningstext" (Neuschwedisches Frequenzwörterbuch basierend auf Zeitungstext²³⁾ Zeitungsmaterial als Objekt seiner Untersuchung, da die Zeitungssprache für die neuschwedische Standardsprache und ihr Schriftbild repräsentativ sei und sie außerdem als ein wichtiges "Massenmedium" verbreitet werde.²⁴⁾

Die Texte wurden aus fünf schwedischen Tageszeitungen exzerpiert, bestehen aus einer Million Wörter und wurden im Jahre 1965 (Januar - Dezember) gesammelt. Es waren 1 387 Artikel von 569 Autoren. Die Artikel repräsentieren folgende Sparten:

1. Allgemeine Reportage
2. Artikel der Kulturseite
3. Artikel der Auslandskorrespondenten

Inhaltlich wird das Material in sechs Gruppen geordnet:

1. Naturwissenschaften mit Erläuterungen
2. Kultur mit Erläuterungen
3. Politik und Gesellschaftsfragen
4. Erwerbsleben und Handelsverkehr
5. Mensch und Milieu ("human interest")
6. Künstlerische Tätigkeiten

Aus praktischen Gründen wurde das gesamte Wortmaterial nur initialalphabetisch und finalalphabetisch sortiert. Die Basis-Vokabular-Liste berücksichtigt nur die Wörter, die eine modifizierte Frequenz von 4,125 und darüber aufweisen.

2.1.3. Englisch

Eines der ältesten Frequenzwörterbücher der englischen Sprache ist "A Measuring Scale for Ability in Spelling" (New York 1915) von P. Ayres mit einem Textumfang von 368 000 Wörtern, der Geschäfts- und Privatbriefen entnommen ist.

R.C. Eldrige, der die Notwendigkeit erkannte, den Auswanderern, die in den um die Niagara-Fälle angesiedelten Fabriken arbeiteten, Englisch zu lehren, zählte in seinem "Six Thousand Common English Words" (Niagara Falls 1911) 58 Kolumnen der vier Buffalo-Zeitungen aus, d.h. einen Umfang von 43 989 Wörtern, mit dem Ziel, einen begrenzten aber konzentrierten Wortschatz des allgemeinen Gebrauchs zu ermitteln.

G. Dewey gab an seiner "Relativ Frequency of English Speech Sounds" (Cambridge 1923) neben der Wortfrequenz auch die Silben- und Lautfrequenz an, da sein Frequenzwörterbuch als Hilfsmittel für die Stenographie bestimmt war. Sein Material von 100 000 Wörtern hatte er dabei hauptsächlich aus Zeitungen und Zeitschriften genommen.

Ausschließlich didaktische und pädagogische Ziele verfolgte E.L. Thorndike in seinen drei Frequenzwörterbüchern mit den Titeln: "Teacher's Word Book", (Columbia University 1921) "The Teacher's Word Book of Twenty Thousand Words" (Columbia University 1931 - 32) und "The Teacher's Word Book of 30 000 Words" (Columbia University 1938). In das erste nahm er, vornehmlich aus der Jugendliteratur exzerpiert, 4 565 000 Wörter auf, von denen er die 10 000 häufigsten ermittelte. Im zweiten vergrößerte er den Umfang der Auswahl auf 10 Millionen Wörter und stellte die 20 000 häufigsten fest und im dritten Band schließlich bei gleichbleibender Auswahl die 30 000 häufigsten. Zudem bestimmte er in den ersten beiden Bänden außer der Wortfrequenz auch die sogenannte "Kreditangabe", die die Anzahl der Quellen, aus denen das betreffende Wort stammt, bezeichnet.

Die Bedeutung der Thorndikeschen Frequenzwörterbücher im Englischen ist etwa mit der des Käding'schen im Deutschen vergleichbar.

N.R. Frensch, C.W. Carter und W. Koenig stützten ihre Untersuchung "The Words and Sounds of Telephone Conversations"²⁵⁾ im Gegensatz zu den damals üblichen Zählungen auf das gesprochene Material, und zwar auf Telephongespräche der Stadt New York. Dabei exzerpierten sie insgesamt 79 390 Wörter.

2.2. Frequenzwörterbücher aus dem Bereich der romanischen Sprachen:

2.2.1. Französisch

Texte aus acht verschiedenen Stilgattungen - Schultexte, philosophische und künstlerische Prosa, Zeitschriften- und Zeitungsartikel, politische Essays, Poesie und Telegramme - legte V.A.C. Henmon seinem "A French Word Book Based on a Count of 400 000 Running Words" (University Wisconsin 1924) zugrunde und erhielt als Resultat eine Wortfrequenzliste der 3 905 häufigsten Wörter im Französischen.

G.E. van der Beke erweiterte in seinem 'French Word Book' (MacMillan 1929) das Henmon'sche Frequenzwörterbuch, indem er ein Wortmaterial von 1 147 748 Wörtern exzerpierte und diese schwerpunktmäßig in drei Abschnitte zusammenfaßte:

1. die 69 häufigsten Wörter von Henmon
2. eine Frequenzliste mit Angaben über die Frequenz und die Anzahl der Wortquellen
3. eine alphabetische Liste mit Angaben über die Quellenanzahl, die jeweilige Wortfrequenz der Texte der herangezogenen Autoren, die Wortfrequenz in Henmons Frequenzwörterbuch und die Frequenz des Gesamtmaterials.

Von primärer Bedeutung in diesem Frequenzwörterbuch ist die Angabe der Quellenanzahl, erst an sekundärer Stelle steht die Frequenz (vgl. die Thorndikesche "Kreditangabe").

2.2.2. Spanisch

Das erste spanische Frequenzwörterbuch 'A Graded Spanish Word Book' von M.A. Buchanan (Toronto 1927) ähnelt methodisch und dem Aufbau nach dem von Thorndike. Buchanan zählte 1 200 000 Wörter aus 40 Quellen und 7 Stil Kategorien.

V. Garcia Hoz stellt in seinem "Vocabulario usual, comun y fundamental (determinación y analisis de sus factores)" (Madrid 1953) den Grundwortschatz des Spanischen mit den 208 häufigsten Wörtern zusammen. Das für dieses Wörterbuch zugrunde gelegte Korpus besteht aus 400 000 Wörtern aus vier Stilgruppen.

Das 'Frequency Dictionary of Spanish Word's von A. Juilland und E. Chang-Rodriguez (The Hague 1964) gehört schon zu den Frequenzwörterbüchern der zweiten Etappe, da hier im Rahmen der quantitativen Analyse drei Koeffizienten: Frequenz (Frequency), Textverteilung (Dispersion) und Wortgebrauch (Usage) ermittelt werden. Es wurden 500 000 Wörter aus 5 Stil Kategorien als Grundlage für dieses Wörterbuch exzerpiert.

2.2.3. Italienisch

U. Bartolini, G. Tagliavini und A. Zampoli stellten unter dem methodologischen Einfluß der Frequenzwörterbücher Juillands das "Lessico di frequenza della lingua italiana contemporanea" (Rom 1971) zusammen. Abweichend von den anderen in dieser Arbeit behandelten Wortfrequenzuntersuchungen hat diese die Stilgattungsauswahl: Drama, Roman, Filmszenario, Publizistik, Grundschullehrbücher.

2.2.4. Rumänisch

Das "Frequency Dictionary of Roumanian Words" von A. Juilland, P.M.H. Edwards und I. Juilland (The Hague 1965) geht mit der Zeitspanne seiner Quellen bis zum Jahre 1890 zurück, um eine Auswahl geeigneter Quellen zu finden.

Das Korpus von 430 000 Wörtern wurden aus fünf Stilgattungen exzerpiert:

1. Drama
2. Prosa
3. Essay
4. Zeitschriften
5. Technische Texte

Hier werden auch die drei Koeffizienten (vgl. Juilland und Chang-Rodriguez 1964) des Rumänischen ermittelt. Laut Haarmann²⁶⁾ sei die Auswahl der Wörter nicht repräsentativ genug, bzw. zu einseitig, da in diesem Frequenzwörterbuch drei Wochentagbezeichnungen fehlten, abgesehen von üblichen Körperteilbezeichnungen oder Obstbezeichnungen.

2.3. Wortfrequenzuntersuchungen aus dem slavischen Sprachbereich

2.3.1. Čechisch

J. Sedláček bearbeitete in seinem bisher unveröffentlichten Manuskript²⁷⁾ "Základní studie k českému tžsnopisu" 100 000 Wörter, speziell für den Bereich der Stenographie, indem er ähnlich Käding die Wörter untransformiert zählte.

J. Korejs und H. Korejsová veröffentlichten im Jahre 1938 ihr "Slovník našich novín",²⁸⁾ dem sie die Zeitung "Národní osvobození" vom 8.1.1928 mit einem Gesamtkorpus von 42 448 Wörtern zugrundelegten.

T. Hamerskýs "Slovník šestiletých dětí" (Gottwaldov 1944) basiert auf einer repräsentativen Auswahl von 109 369 Wörtern aus der Kindersprache, die stenographisch aufgenommen wurden. Analog dazu ging M. Bartušková in ihrem "Slovník šestiletých dětí" (Praha 1945) von 51 981 Wörtern aus. A. Hobzík und F. Šebek bestimmten die Frequenz der Wörter von fünf tschechischen ABC-Büchern.²⁹⁾

M. Matula, M. Matoušek und V. Freml zählten speziell für den stenographischen Bereich ein Korpus von 60 000 Wörtern aus, wobei sie als Quellenmaterial die stenographischen Aufzeichnungen der Parlaments-sitzungen von 1951 bis 1952 benutzten. 1957 erweiterten Matula und Matoušek das Korpus um 12 000 Wörter aus der Geschäftskorrespondenz. Als Ergebnis dieser Auszählungen veröffentlichte Matula schließlich die "Teorie a praxe těsnopisu", in der er auch die von Freml aufgestellten Statistiken berücksichtigte.

Das "Frekvence slov, slovních druhů a tvarů v českém jazyce" (Praha 1961) ist eine Kollektivarbeit von J. Jelínek, J. Bečka und M. Těšitelová. Gezählt und sortiert wurden 1 623 527 Wörter aus 75 Quellen, klassifiziert nach den Genres:

1. Belletristik
2. Poesie
3. Jugendliteratur
4. Drama

5. Fachliteratur
6. Publizistik
7. Wissenschaftliche Literatur
8. Gesprochene Sprache

Das Material stammt aus der Zeitspanne zwischen 1926 und 1950. In der ersten Liste sind die 1 000 häufigsten Wörter nach abnehmender Frequenz eingeordnet. Das Wort ist durch drei numerische Angaben charakterisiert, z.B. každý 2254 - 8 - 75, wobei die erste Zahl die absolute Frequenz, die zweite die Anzahl der Stilgruppen, die dritte die Anzahl der Quellen angibt, in denen das betreffende Wort vorkommt. Die zweite Liste vermittelt die gleiche Information über das Wort wie die erste, ist jedoch alphabetisch geordnet.

2.3.2. Slovakisch

Die ersten Wortfrequenzuntersuchungen im Slovakischen sind von Mistrík in den Jahren 1952 - 1954 durchgeführt worden. Ein Teil seiner Analyse - (Wörter mit der Frequenz 100 und mehr) veröffentlichte er in der Zeitschrift 'Slovenský stenograf'³⁰⁾, den größten Teil in dem Buch 'Stenografia'³¹⁾. Ausgehend von einem Material mit 100 000 Wörtern aus 10 "stilistisch-thematischen"³²⁾ Gattungen, listete er Wörter in der Grundform mit der Frequenz 25 und darüber auf, sowie ihre modifizierte Formen mit der Frequenz 5 und mehr.

Das Material seines späteren "Frekvencia slov v slovenčine" (Bratislava 1969) stammt aus 60 Texten mit insgesamt 1 000 000 Wörtern, die einen Zeitraum von 40 Jahren umfassen, und wurde in fünf Stilgruppen eingeteilt:

1. Dialoge, d.h. Drama
2. Künstlerische Prosa
3. Poesie
4. Journalistik
5. Wissenschaftliche Prosa

In die numerisch geordnete Frequenzliste sind 9 568 Wörter nach abnehmender Frequenz eingeordnet; im Gegensatz zu den bisherigen Untersuchungen wurden Rang und Ordnungszahl des Wortes hier getrennt aufgeführt. Jeweils 10 Wörter werden zu einer Gruppe zusammengefaßt. Hinter jeder Gruppe steht die kumulative Summen- und Prozentangabe über die relative Frequenz.

Die alphabetische Liste gibt die 21 823 häufigsten Wörter. Den Kern des slowakischen Wortfundaments, d.h. des lexikalischen Standards, der aufgrund der rein formalen Kriterien zusammengestellt wurde, setzte er mit 2 573 Wörtern fest.

"Zum Kern des Wortfundaments werden in der Regel alle primären, fundamentalen, einheimischen und eingebürgerten Wörter gerechnet, die im gebräuchlichen Verständigungsverkehr unvermeidbar sind."³³⁾

Den Abschluß seines slowakischen Frequenzwörterbuches bilden Frequenzlisten mit den 500 häufigsten Wörtern des Čechischen, Russischen, Deutschen, Englischen, Französischen, Spanischen und Romanischen, entnommen aus einschlägigen Untersuchungen.

2.3.3. Russisch

Schon im Jahre 1933 ging man in der Sowjetunion daran, den Wortschatz eines Dichters mit quantitativen Methoden zu analysieren. So untersuchte u.a. K.M. Frumkina in "Materialy k častotnomu slovarju jazyka Puškina"³⁴⁾ den Wortschatz aller schriftlichen Äußerungen Puškins, wobei sie, bezogen auf ein Korpus von 544 777 Wörtern, zu dem Ergebnis gelangte, daß die 2 000 häufigsten Wörter jeweils 80 % eines beliebigen Puškin-Textes decken.

Das "Russian Word Count" von H.H. Josselson³⁵⁾ ist das erste Frequenzwörterbuch, das mit Hilfe eines Computers erstellt wurde. Die zeitliche Streuung der Texte beträgt 120 Jahre. Dem Genre nach teilt sich das Gesamtmaterial wie folgt:

1. Drama
2. Literaturkritik
3. Journalistik
4. Belletristik

Die Anzahl der gezählten Wörter beträgt 1 000 000. Das Wörterbuch enthält zwei Register mit jeweils 5 230 Wörtern. Das erste ist der Frequenz nach in sechs Subregister unterteilt, das zweite alphabetisch geordnet. Die Besonderheit dieser Untersuchung liegt darin, daß bei den Frequenzangaben Rücksicht auf die Chronologie der Texte genommen wird.

Das "Častotnyj slovar' sovremennogo russkogo jazyka" von E. Štejnfel'dt mit dem Untertitel "Die 2 500 am häufigsten gebrauchten Wörter"³⁶⁾ für estnische Schulen basiert auf einem Korpus von 400 000 Wörtern.

Betrachtet man jedoch das zugrunde gelegte Material, so erhebt sich allerdings die Frage, ob eine solchegattungsspezifische Textauswahl die - laut Štejnfeldt - "gegenwärtige Literatursprache des Russischen" reflektieren kann.

Hier war es vor allem L.N. Zazorina, in deren "Častotnyj slovar' russkogo jazyka" (Moskva 1977) das moderne Russisch umfassend 1 056 382 Wörter aus vier stilistischen Gattungen

1. Wissenschafts- und Fachtexte
2. Künstlerische Prosa
3. Drama
4. Journalistik

Berücksichtigung fand.

2.3.4. Polnisch

Im Polnischen wären die Arbeiten "Frekwencja wyrazów w prasie" (Krakow 1972) von W. Pisarek und "Słownictwo współczesnej publicystyki polskiej, Listy frekwencyjne" (Warszawa 1971) von A. Lewicki, W. Masłowski und J. Sambor zu nennen.

2.3.5. Altkirchenslavisch

Einen wertvollen Dienst für die altslavische Philologie hat B. Anemüller³⁷⁾ mit seiner Untersuchung "Maschinelle Lexikographie" geleistet. Er wendet die Methoden der linguistischen Datenverarbeitung auf altbulgarisches Material an, und zwar eine altbulgarische Übersetzung aus dem Griechischen der spätantiken-frühbyzantinischen Literaturperiode mit dem Titel "Fragen und Antworten" des Pseudo-Kaisarios.

Als konkretes Ergebnis seiner Arbeit entstand schließlich ein lemmatisierter Wortformenindex dazu. Bei dieser Untersuchung handelt es sich nicht um eine kritische Textedition. Vielmehr wollte der Verfasser bisher unbeachtetes oder schwer zugängliches Sprachmaterial unter weitgehender Wahrung der Textautonomie transparent machen und als Forschungsgegenstand möglichst rasch zur Verfügung stellen.³⁸⁾ Eine differenzierte grammatische, syntaktische oder semantische Klassifizierung der Textelemente klammert er aus. Der generelle Vorteil einer maschinellen Bearbeitung des Materials lag im ganz konkreten Falle in der Schaffung eines 'Computer-verfügbaren' Textarchivs, in der beliebigen Wiederholbarkeits sämtlicher Vorgänge, in den verschiedenen Sortiermöglichkeiten des Materials und schließlich in der beliebigen Ausbaufähigkeit des Textmaterials bei geeigneter Speicherung. Die Aufstellung einer Wortfrequenzliste des Materials ließe sich auch mit Leichtigkeit durchführen.

Das "Elektronische Archiv des altslavischen Wortschatzes" (L'Archivo elettronico del lessico paleoslavo AELP) am Institut für slavische Philologie der Universität in Padua hat unter Leitung von N. Radovich³⁹⁾ hauptsächlich das Textmaterial des "Handwörterbuches zu den altkirchenslavischen Texten" von L. Sadnik und R. Aitzetmüller auf Magnetbändern gespeichert und per Computer die Zahl der Wörter und ihre prozentuale Verteilung in den Denkmälern errechnet und ermittelt. Radovich will durch seine rein arithmetische Rechnung einen Überblick über den Hauptwortschatz der altkirchenslavischen Texte und über die Schichten des Wortschatzes geben, die typisch für literarische Stilarten, Schriftzentren oder religiöse Strömungen sind. Das Verdienst seiner Arbeit besteht darin, den gemeinsamen Wortschatz

verschiedener altkirchenslavischen Texte ermittelt zu haben, allerdings leider ohne eine Zusammenstellung von Wortfrequenzlisten.

2.4. Wortfrequenzuntersuchungen aus dem finno-ugrischen Sprachbereich

2.4.1. Finnisch

Das erste umfassende Frequenzwörterbuch der finnischen Sprache ist das im Jahre 1979 erschienene "Suomen kielen taajuussanasto" von Saukkonen, Haipus, Niemikorpi und Sulkala.⁴⁰⁾ Das Grundkorpus bilden 408 301 Wörter, die aus folgenden Quellen exzerpiert wurden: aus für den Rundfunk bestimmten Sprechtexten, aus der wissenschaftlichen Literatur, aus Zeitungen und Zeitschriften und aus der schöngeistigen Literatur. Für die endgültige Auswahl wurden insgesamt 5 800 Textabschnitte mit mindestens 60 fortlaufenden Wörter ausgelost. Als Ergebnis entstand eine Wortfrequenzliste, in der die Wörter mit der Frequenz 2 und mehr vorkommen, und eine alphabetische Liste mit den gleichen Daten.

Das 1970 zusammengestellte "Opettajien sanastokirja" (Lehrerwörterbuch)⁴¹⁾ dient ausschließlich pädagogischen Zwecken. Es macht Aussagen über die Einflüsse verschiedener Faktoren, wie Alter, Geschlecht, sozialer Hintergrund auf die Wortfrequenz im schriftlichen Sprachgebrauch von Schülern der dritten und vierten Grundschulklassen. Als Material für dieses Wörterbuch dienten Aufsätze von 8 708 Schülern. Insgesamt wurden 840 000 Wörter gezählt, die hier nach abnehmender Frequenz sortiert und nach dem Alphabet geordnet wurden.

2.4.2. Estnisch

Im Estnischen basiert die bisher einzige Frequenzliste der Wortformen und die daraus entstandene Lexemliste auf dem Material der zeitgenössischen schöngeistigen Prosa, bestehend aus einem Korpus von 100 000 Wörtern. Sie wurde in Teamarbeit von Kaasik, Tuldava, Villup und Ääremaa verfaßt und im Jahre 1977 in einer linguistischen Zeitschrift der Tartuer Universität veröffentlicht.⁴²⁾

2.4.3. Ungarisch

Für den stenographischen Gebrauch sortierte Z. Nemes in seiner "Neuen ungarischen Wörterstatistik"⁴³⁾ aus einem Korpus von 397 956 Wörtern 31 078 Lexeme heraus, von denen 15 171 (= 48,82 %) nur ein einziges Mal vorkamen, sowie 16 571 verschiedene wortbildende Morpheme. An einem neuen Frequenzwörterbuch der ungarischen Sprache wird unter der Leitung von J. Kelemen gearbeitet.

III. LEXIKOSTATISTISCHE VERGLEICHE IN DER
ČECHISCHEN UND FINNISCHEN SPRACHE

1. DISTRIBUTION DER LEXIKALISCHEN EINHEITEN

1.1. Klassifizierung nach P. Guiraud

Wenn man die Wörter zusammenhängender Texte nach der Frequenz ihres Vorkommens aufreihet, fällt in der Aufreihung ein bestimmtes System auf.

Die am häufigsten vorkommenden Wörter, die die ersten Stellen in den Frequenzwörterbüchern besetzen, haben auch ihre bestimmten lexikalisch-grammatikalischen Eigenschaften: Sie sind vor allem Funktionswörter, und erst nach ihnen kommen die autosemantischen Wörter.

Unter diesem Gesichtspunkt teilt P. Guiraud¹⁾ die Wörter nach ihrer Frequenz in vier Gruppen ein:

1. dominierende, d.h. strukturelle und grammatische Wörter ("sémantiquement vides"), die ungefähr die Hälfte eines zusammenhängenden Textes ausmachen. Im Französischen gibt es ca. 100 dieser Wörter.
2. thematische Wörter. Ihre Anzahl beträgt ca. 50 und macht 9 % eines zusammenhängenden Textes aus.
3. Basiswörter, d.h. Wörter, die zum zentralen Wortschatz gehören. Davon gibt es ca. 4 000.
4. charakterisierende Wörter ("mots de caractérisation"); etwa 20 000 an der Zahl.

1.2. Wortschatzgliederung nach J. Mistrík

Das Ausgangsmaterial für den Vergleich des Systems der sogenannten vertikalen Wortdistribution entnahm Mistrík fünf Frequenzwörterbüchern (slowakisch, tschechisch, russisch, deutsch und spanisch). Für die vertikale Verteilung galten die Grundsätze der Frequenz und der Semantik. Der Vergleich bezieht sich hierbei auf 110 dominierende lexikalische Einheiten innerhalb einer Wortfrequenzliste. Diese machen in allen Sprachen ca. 50 % eines beliebigen Textes aus, während die restlichen Wörter gemäß ihrer Semantik gewählt und verteilt werden.

Diese dominierenden Wörter versteht Mistrík als parametrische Mengen, mit Hilfe derer sich der Wortschatz verschiedener Autoren untersuchen läßt.

Der Wortschatz wird folgendermaßen gegliedert:²⁾

1. Konstruktionswörter 'K', zu denen grammatische Wörter wie Konjunktionen, Artikel, Hilfsverben etc. gehören. Sie nehmen in den Frequenzwörterbüchern die Plätze 1 - 20 ein.
2. Subjektive, von der Situation determinierte Wörter 'S' (auch subjektiv-situativ genannt), zu denen die Modalverben, Personalpronomina, einige Partikel und Interjektionen gehören. Sie nehmen die Plätze 21 - 70 ein.
3. Gnomische Wörter 'G', zu denen Ausdrücke mit allgemeiner, objektiver Bedeutung wie Numeralia und Nomina

gezählt werden. Diese nehmen die Plätze 71 - 110 ein.

4. Thematische Wörter 'T', alle diejenigen Wörter, die in den oben genannten Schichten nicht vorkommen. Es handelt sich dabei vorwiegend um Substantive, Adjektive und Verben mit spezieller und terminologischer Bedeutung, die die Plätze 111 ff. einnehmen.

Diese Einteilung in vier Gruppen erfolgt primär nach dem Aspekt der Frequenz und erst sekundär nach dem der Semantik, denn erstere erlaubt eine exakte Einteilung in Schichten, während im zweiten Falle die Grenzen fließend sind.

Das Vorgehen ist durch Anwendung exakter Methoden untermauert, jedoch gilt es dabei in Erinnerung zu rufen, daß eine quantitative Information immer nur das Signal eines inhaltlichen Phänomens ist.³⁾

Den Kern der ersten Schicht bilden grammatisch-konstruktive Elemente: Präpositionen und Konjunktionen. Sie verweisen auf die grammatischen Beziehungen. Die Wörter 'býti', 'mīti', 'který', 'ten' üben jedoch der Statistik nach vorwiegend eine grammatische Funktion aus.

Den Kern der zweiten Schicht bilden die Wörter, die in der unmittelbaren Kommunikation ihren Wert erhalten. Mit ihrer Hilfe wird eine bestimmte Haltung, (Partikeln, Modalverben etc.), bzw. eine Einstellung zum Millieu (deiktische Pronomina, Adverbia) oder zum Adressaten (Personal- und Possessivpronomina) ausgedrückt.

Zum Kern der dritten Schicht gehören die autosemantischen Wörter, mit denen allgemein und nicht-terminologisch die Phänomene (Gegenstände, Eigenschaften, Handlungen, Umstände) benannt werden. Zu ihnen gehören die Basiswörter 'lidé', 'ruka', 'doba', 'čas', 'věc', 'hlava', 'mysliti', 'přijíti'.

Bei der Verteilung der 110 dominierenden, häufigsten Wörter auf die drei Schichten sind die Grenzen der Zonen semantisch nicht eindeutig und klar.

In der Gruppe der Konstruktionswörter sind auch subjektive, von der Situation determinierte Wörter, manchmal sogar gnomische Wörter. Daher sind in jeder Schicht mindestens zwei Ebenen vorhanden, klassifiziert nach Frequenz bzw. Semantik, die sich gegenseitig überschneiden.

Die nach dem Aspekt der Frequenz geordneten werden mit 'K', 'S', 'G' bezeichnet, während die nach dem Aspekt der Semantik geordneten 'k', 's', 'g' genannt werden.

Die von Mistrík aufgestellte These, daß in jeder Sprache an erster Stelle die 'K'-Wörter stehen, an zweiter die 'S'- und an dritter die 'G'-Wörter, hat sich in den von ihm verglichenen Sprachen - alle zur indogermanischen Sprachgruppe gehörend - bewährt. Daß dies auch auf das Finnische als nicht-indogermanische Sprache zutrifft, wird im Rahmen dieser Arbeit noch aufgezeigt werden.

1.3. Vertikale Wortdistribution des Čechischen und Finnischen

1.3.1. Wortfrequenzlisten

Die Čechische Wortfrequenzliste ist aus dem Čechischen Frequenzwörterbuch "Frekvence slov, slovních druhů a tvarů v českém jazyce" von J. Jelínek, J. Bečka und M. Těšitelová und die des Finnischen aus dem finnischen Frequenzwörterbuch "Suomen kielen taajuussanasto" von P. Saukkonen, M. Haipus, A. Niemikorpi und H. Sulkala entnommen.

In der Liste sind die 110 häufigsten Wörter der beiden Sprachen aufgeführt.

Die Gruppeneinteilung in 'K', 'S', 'G' und die eventuelle Kollision mit 'k', 's', 'g' wird dabei auch angezeigt.

Die horizontalen Linien markieren die Grenzen zwischen den Gruppen 'K', 'S' und 'G'.

Čechisch

Finnisch

a (K)
býti (K)
ten (K)
v(e) (K)
on (K,s)
na (K)
že (K)
s(e) (K)
z(e) (K)
10 který (K)

olla (K)
ja (K)
se (K,s)
ei (K,s)
joka (K)
että (K)
tämä (K,s)
hän (K,s)
voida (K,s)
10 saada (K,s)

mīti (se) (K)	mutta (K)
jā (K,s)	niin (K)
k, ke, ku (K)	kuin (K)
do (K)	ne (K,s)
i (K)	kun (K)
ale (K)	tulla (K,g)
svūj (K,s)	myös (K)
jako (K)	tai (K)
o (K)	kaikki (K,s)
20 tak (K)	20 aika (K,g)
co (S,k)	me (S)
za (S,k)	suuri (S,g)
se (S)	vuosi (S,g)
tento (S)	toinen (S,g)
jen (S)	muu (S)
po (S,k)	minä (S)
aby (S,k)	antaa (S,g)
jak (S,k)	pitää (S,g)
30 jenž (S,k)	30 tehdä (S,g)
když (S,k)	jo (S)
všechn (S)	he (S)
už (S)	vain (S)
pro (S,k)	nyt (S)
my (S)	sekä (S,k)
ty (S)	sanoa (S,g)
ještě (S)	kuitenkin (S)
u (S,k)	nämä (S)
tu (S)	käyttää (S,g)
40 museti (S)	40 jos (S,k)
náš (S)	asia (S,g)
věděti (S,k)	maa (S,g)
od(e) (S,k)	sitten (S,k)
takê/y (S,k)	ottaa (S,g)

vy (S)	mukaan (S,k)
chtīti (S)	ihminen (S,g)
nebo (S,k)	vielä (S)
pān (S,g)	osa (S,g)
ani (S,k)	oma (S)
50 jīti (S,g)	50 uusi (S,g)
až (S)	päivä (S,g)
řici (S,g)	sama (S)
jeden (S,g)	itse (S)
viděti (S,g)	suomi (S,g)
při (S,k)	hyvä (S,g)
než (S,k)	suorittaa (S,g)
dāti (S,g)	moni (S,g)
celý (S)	kaksi (S,g)
mŭj (S)	pieni (S,g)
60 život (S,g)	60 jokin (S,k)
pak (S)	sellainen (S)
-li(ž) (S,k)	kuulua (S,g)
člověk (S,g)	sinä (S)
však (S,k)	hyvin (S,k)
již (S)	ensimmäinen (S,g)
jiný (S)	työ (S,g)
sām (S)	yksi (S,g)
před(e)	mennä (S,g)
tam (S)	nähdä (S,g)
70 takový (S,k)	70 jälkeen (S,k)
veliký (G)	mies (G)
práce (G)	tapa (G)
ruka (G)	alkaa (G)
nový (G)	paljon (G,s)
nic (G,s)	tässä (G,s)
její (G,s)	näin (G,s)
den (G)	tapahtua (G)

ne (G,s)
každý (G,s)
80 kde (G,s)

zem(ě) (G)
lidě (G)
proto (G,k)
dva (G)
teda/y (G,k)
doba (G)
ted' (G,s)
mezi (G,k)
nad(e) (G,k)
90 přec(e) (G,s)

něco (G,s)
druhý (G)
hlava (G)
přijíti (G)
oko (G)
zas(e) (G,s)
víc(e) (G,s)
mysliti (G)
pod(e) (G,k)
100 bez(e) (G,k)

snad (G,s)
dnes (G,s)
kdo (G,s)
cesta (G)
chvíle (G)
starý (G)
svět (G)
proti (G,k)
čas (G)
110 věc (G)

eräs (G)
vaan (G,s)
80 mainita (G)

vaikka (G,k)
eri (G,s)
esittää (G)
koko (G,s)
jumala (G)
kysymys (G)
käydä (G)
lapsi (G)
ennen (G,k)
90 jäädä (G)

mieli (G)
kanssa (G,k)
koska (G,k)
siis (G,k)
siellä (G,k)
saattaa (G)
aina (G,k)
katsoa (G)
joutua (G)
100 tietää (G)

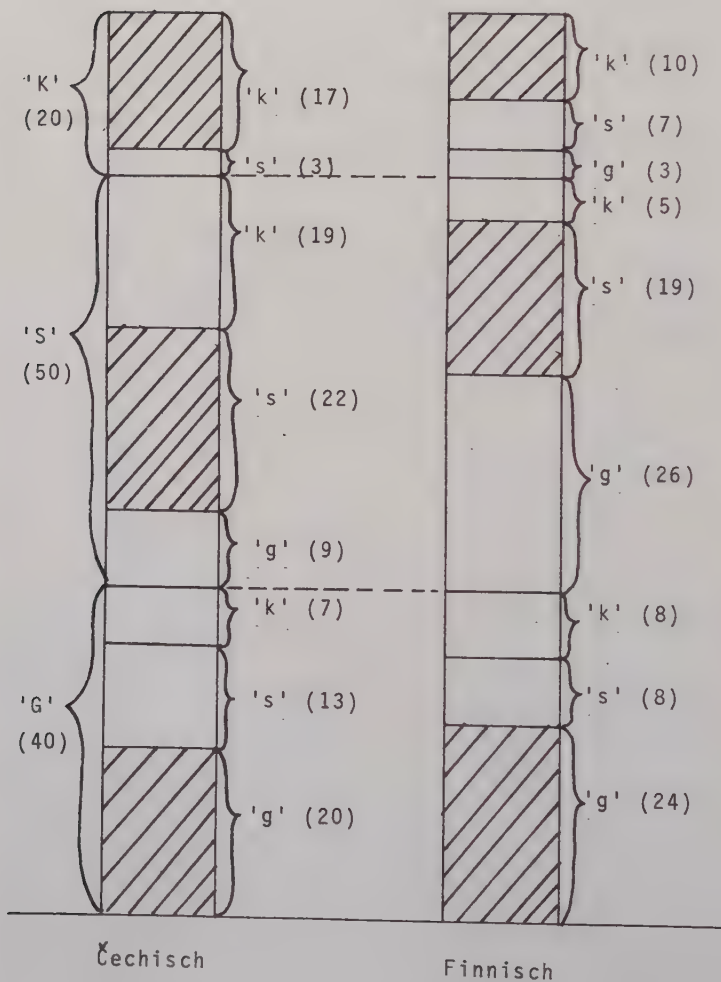
sillä (G,k)
vanha (G)
silloin (G,s)
tuntea (G)
nuori (G)
lähteä (G)
päästä (G)
puoli (G)
tarvita (G)
110 juuri (G,k)

Es lassen sich folgende Überschneidungen beobachten:

In der Menge 'K' im Čechischen sind außer 17 'K'-Elementen 3 Elemente, im Finnischen, abgesehen von 10 'K'-Elementen, 7 Elemente, die semantisch zur Menge 'S' gehören, und im Finnischen sogar 3 zur Menge 'G' gehörende Elemente. In der Menge 'S' sind im Čechischen außer 22 'S'-Elementen 19 zur Menge 'K' gehörende Elemente, 9 zur Menge 'G' gehörende Elemente, in der Menge 'S' im Finnischen wiederum befinden sich außer 19 'S'-Elementen 5 semantisch zur 'K'-Menge gehörende und 26 zur 'G'-Menge gehörende Elemente.

In der Menge 'G' im Čechischen kommen außer 20 'G'-Elementen 7 Elemente von 'K' und 13 von 'S' vor, während das Finnische außer 24 'G'-Elementen 8 von 'K' und 8 von 'S' enthält.

Die graphische Darstellung der Frequenz-Semantik-Äquivalenz der Mengen 'K', 'S', 'G' und 'k', 's', 'g' im Čechischen und Finnischen:



Numerische Darstellung

Gliederung nach Frequenz	Menge 'K'			Menge 'S'			Menge 'G'		
Semantische Gliederung	'k'	's'	'g'	'k'	's'	'g'	'k'	's'	'g'
Čechisch	17	3	ø	19	22	9	7	13	20
Finnisch	10	7	5	5	19	26	8	8	24

Die Symbole und ihre Bedeutung in der graphischen Darstellung:

'k' = Konstruktionswörter

's' = Subjektive, von der Situation determinierte Wörter

'g' = Gnomische Wörter

'K' = Menge der dominierenden Wörter, benannt nach der Menge 'k' der Konstruktionswörter

'S' = Menge der Wörter, die nach der Menge 's' der subjektiven Wörter bezeichnet sind

'G' = Menge der Wörter, die nach der Menge 'g' der gnomischen Wörter benannt sind

Als vierte Menge entsteht durch die vertikale Gliederung des Wortschatzes die Menge 'T', in der thematische und terminologische Wörter vorkommen.

Die Harmonie der äquivalenten Ebenen, Frequenz und Semantik, kann dadurch gestört werden, daß eine Menge 'K', 'S' oder 'G' unter dem Einfluß fremder Elemente einen anderen semantischen Charakter bekommt als die parame. trische Konfiguration erwarten läßt.

Wenn beispielsweise Elemente aus anderen Mengen in die Menge 'K' gelangen, führt dieses nicht zum Schwund eines Elementes, sondern zu seiner Verschiebung nach hinten. Dies gilt analog auch für die anderen Mengen.

Mit Hilfe mathematischer Terminologie und Symbole läßt sich dies, wie folgt, darstellen:

Dabei bedeutet

r	=	Rang eines Wortes im Frequenzwörterbuch
K	=	Menge der Wörter mit $r = 1 - 20$
S	=	Menge der Wörter mit $r = 21 - 70$
G	=	Menge der Wörter mit $r = 71 - 110$
T	=	Menge der Wörter mit $r = 111$
k	=	Menge aller Konstruktionswörter
s	=	Menge aller subjektiven, von der Situation determinierten Wörter
g	=	Menge aller gnomischen Wörter
t	=	Menge aller thematischen Wörter

Es sei K_k eine Untergruppe in der Menge der Konstruktionswörter, die in die Menge $r_{1-20}(K)$ gehören, K_s

eine Untergruppe der subjektiven Wörter, die der Menge $r_{1-20}(K)$ angehören, und S_k eine Untergruppe der Konstruktionswörter, die der Menge $r_{21-70}(S)$ zugeordnet sind. Ähnliches gelte für S_s, S_g, S_t und G_k, G_s, G_t .

Das bedeutet:

$$K_k \cong K, K_s \cong K$$

$$S_k \cong S, S_s \cong S, S_g \cong S$$

$$G_k \cong G, G_s \cong G, G_g \cong G, G_t \cong G$$

$$T_k \cong T, T_s \cong T, T_g \cong T, T_t \cong T$$

Die 110 häufigsten tschechischen Wörter ergeben folgende Zeichendefinition:

$$K_s < K_k \cong K$$

$$S_k < S_s \cong S$$

$$S_g < S_s \cong S$$

$$G_k < G_g \cong G$$

$$G_s < G_g \cong G$$

Das entsprechende im Finnischen:

$$K_s < K_k \cong K$$

$$K_g < K_k \cong K$$

$$S_k < S_s \cong S$$

$$S_g > S_s \leq S$$

$$G_k < G_g \leq G$$

$$G_s < G_g \leq G$$

Abweichend von Čechischen und zusätzlich hat das Finnische die Definition $K_g < K_k \leq K$

und nur abweichend $S_g > S_s \leq S$.

Dies wird einerseits durch den Mangel an Präpositionen, andererseits durch die dadurch entstandene Möglichkeit der autosemantischen Wörter, unter die häufigsten Wörter zu dringen, verursacht.

1.3.2. Vertikale Wortdistribution Čechisch/Slovakisch Finnisch/Estnisch

Das Slovakische ist dem Čechischen eng verwandt, etwa in einem ähnlichen Verhältnis wie das Estnische mit dem Finnischen.

Vergleicht man nun im Čechischen und Slovakischen die jeweils 110 dominierenden Wörter miteinander, so zeigt sich bei der vertikalen Wortdistribution eine auffällige Schwankung in den Konstruktions- und gnomischen Wörtern.

Im Finnischen und Estnischen hingegen divergieren nur die Konstruktionswörter.

Dies läßt sich aus der graphischen und numerischen Darstellung der Frequenz-Semantik-Äquivalenz der

Mengen 'K', 'S', 'G' und 'k', 's', 'g' auf den Seiten 64 und 65 entnehmen.

Wenn man den Anteil der Wörter, die zu den sogenannten Untergruppen K_k , S_s , G_g gehören, bei denen die Einteilung nach Frequenz mit der der Semantik übereinstimmt, beider nahverwandten Sprachen jeweils miteinander vergleicht, ist der Anteil der tschechischen Wörter annähernd übereinstimmend mit dem der slowakischen, während zwischen den finnischen und estnischen größere Abweichungen sind.

tschechisch/slowakisch

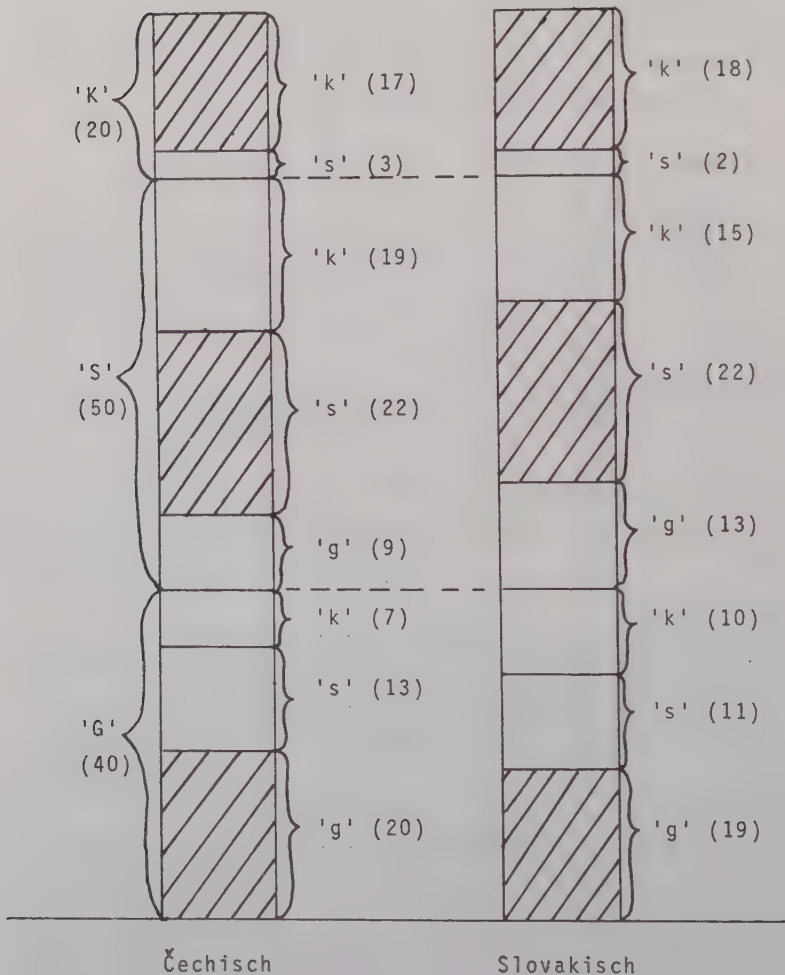
finnisch/estnisch

K_k	17:18	8:12
-------	-------	------

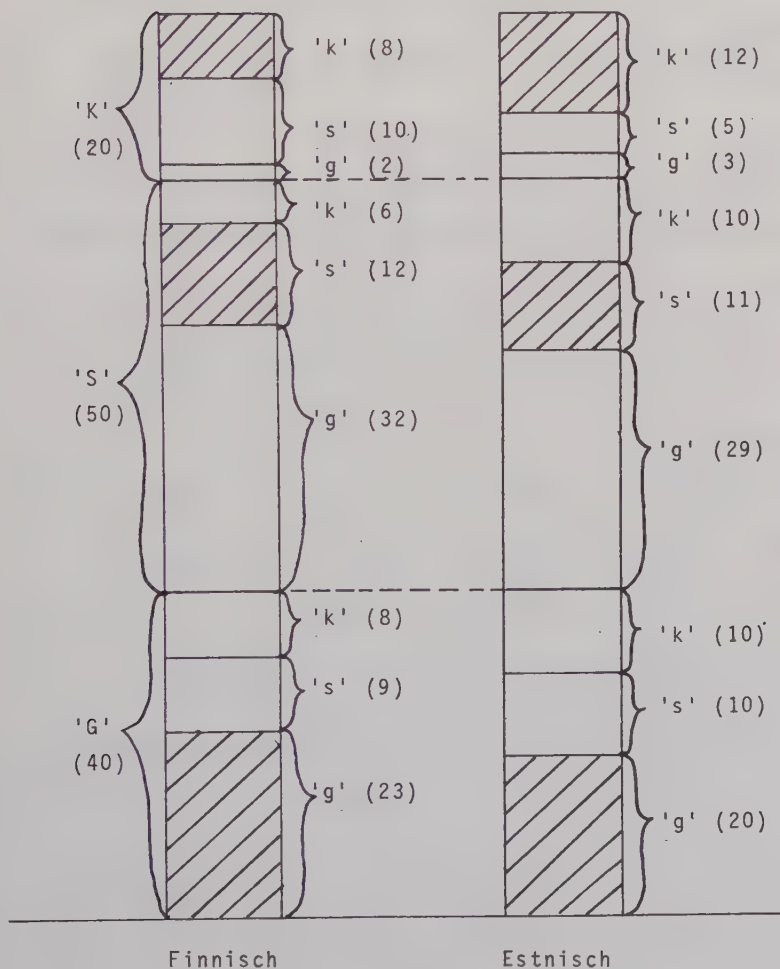
S_s	22:22	12:11
-------	-------	-------

G_g	20:19	23:20
-------	-------	-------

Die graphische Darstellung der Frequenz-Semantik-Äquivalenz der Mengen 'K', 'S', 'G' und 'k', 's', 'g' im Čechisch und Slovakisch:



Die graphische Darstellung der Frequenz-Semantik-Äquivalenz der Mengen 'K', 'S', 'G' und 'k', 's', 'g' im Finnischen und Estnischen:



Die numerische Darstellung der Čechischen, Slovaki-
schen, Finnischen und Estnischen ergibt folgendes
Bild:

Gliederung nach Frequenz	Menge 'K'			Menge 'S'			Menge 'G'		
	'k'	's'	'g'	'k'	's'	'g'	'k'	's'	'g'
Čechisch	17	3	Ø	19	22	9	7	13	20
Slovakisch	18	2	Ø	15	22	13	10	11	19
Finnisch	8	10	2	6	12	32	8	9	23
Estnisch	12	5	3	10	11	29	10	10	20

1.3.3. Formale Differenzen

Formal unterscheiden sich die Wörter der einzelnen Mengen 'K', 'S', 'G' voneinander. Der durchschnittlichen Silbenlänge im Čechischen von 1,25 in der Menge 'K' entspricht eine Silbenlänge von 1,75 im Finnischen.

Die Wörter der Menge 'S' im Čechischen weisen eine Silbenlänge von 1,50 auf, während dieselben im Finnischen zweisilbig sind.

Die zur Menge 'G' gehörenden Wörter haben im Čechischen eine Silbenlänge von 1,70 und im Finnischen von 2,20.

Dasselbe tabellarisch dargestellt (wobei auch die entsprechenden Werte des Slovakischen und Estnischen aufgeführt werden):

Menge	'K'	'S'	'G'
Čechisch	1,25	1,50	1,70
Slovakisch	1,15	1,25	1,77
Finnisch	1,75	2,00	2,20
Estnisch	1,75	1,92	1,75

Die durch die vertikale Gliederung des Wortschatzes entstehende Menge 'T' wird von thematischen und terminologischen Wörtern gebildet. Diese Wörter sind alle deklinierbar und haben im Durchschnitt mehr als zwei Silben.⁴⁾

IV. GRAMMATISCH - STATISTISCHE VERGLEICHE
IN DER ČECHISCHEN UND FINNISCHEN
SPRACHE

1. WORTKLASSENDISTRIBUTION

Das Ziel der grammatischen Statistik unterscheidet sich im allgemeinen von dem der Lexikostatistik: Mit Hilfe der statistischen Methoden teilt man im Lexikon die lexikalischen Einheiten in verschiedene Kategorien ein; in der Grammatik handelt es sich vor allem um eine Hierarchie der einzelnen Einheiten, der Kategorien. Die statistischen Methoden können nicht, laut Těšitelová, eine vollständige Beschreibung der grammatischen Kategorien gewährleisten, aber umgekehrt sei jede Beschreibung der Grammatik und ihrer Kategorien ohne quantitative Daten zweifellos unvollkommen.¹⁾

Welchen Stellenwert hinsichtlich der Quantität die einzelnen Wortklassen im Čechischen und Finnischen einnehmen, läßt sich aus den folgenden Listen entnehmen. In der Rangordnung (= R) ihrer Frequenz, einmal nach der textualen (= T), einmal nach der lexikalen (= L) Frequenz, stehen die Wortklassen aufgeführt und bilden eine Hierarchie, die im Čechischen bei den drei häufigsten Wortklassen unverändert, im Finnischen leicht verändert bleibt.

Čechisch:

R Wortklasse	% T	R Wortklasse	% L
1. Substantive	27,8	1. Substantive	38,9
2. Verben	18,2	2. Verben	27,1
3. Adjektive	11,2	3. Adjektive	21,0
4. Pronomina	10,9	4. Adverbien	9,0
5. Adverbien	10,3	5. Pronomina	1,0
6. Präpositionen	10,1	6. Konjunktionen	0,9
7. Konjunktionen	9,8	7. Präpositionen	0,7

Finnisch:

R Wortklasse	% T	R Wortklasse	% L
1. Substantive	35,8	1. Substantive	75,3
2. Verben	24,3	2. Adjektive	10,0
3. Adjektive	9,2	3. Verben	8,2
4. Adverbien	8,7	4. Adverbien	1,9
5. Pronomina	8,4	5. Präpositionen	0,2
6. Konjunktionen	7,5	6. Pronomina	0,1
7. Präpositionen	1,0	7. Konjunktionen	0,1

In den auf den Seiten 72 und 73 stehenden Tabellen (Tab.1 und 2) ist die textuale und lexikale Verteilung, d.h. die Distribution der Lexeme und Textwörter, beider Sprachen nach ihren grammatischen Funktionen angegeben.

Die Lexeme mit der anteilmäßig stärksten Vertretung innerhalb der häufigsten Wörter sind im Čechischen wie im Finnischen die Substantive, an der zweiten Stelle stehen im Čechischen die Verben, im Finnischen die Adjektive, an der dritten Stelle Adjektive im Čechischen und Verben im Finnischen. Während die anderen Wortklassen mit Bezug auf das lexikalische Gesamtinventar einen kleinen, beispielsweise Adverbien im Čechischen, bzw. minimalen Anteil (z.B. Konjunktionen) ausmachen, treten sie hinsichtlich ihrer textualen Frequenz deutlich in den Vordergrund.

Wenn man nun als ein weiteres Kriterium die absolute Frequenz in Betracht zieht und die Angaben nach Wortklassen differenziert, so ergibt sich eine stärkere anteilmäßige Vertretung der Adverbien, Präpositionen, Pronomina und Konjunktionen im Čechischen, und im

Finnischen der Adverbien, Pronomina und der Konjunktionen.

Durch die Graphiken auf den Seiten 74 und 75 werden die Proportionen der Lexeme im Verhältnis zu den Textwörtern veranschaulicht.

Erklärungen zu den Tabellen und den Graphiken:

In den Tabellen und Graphiken wird als achte Wortklasse 'Sonstige' aufgeführt. Es sind Interjektionen, Numeralia, die aus dem finnischen Wortmaterial²⁾ von vornherein aussortiert wurden, oder solche, die nicht eindeutig in eine Wortklasse einzuordnen waren, wie z.B. 'tosi' und 'paitsi'.

Aus Homogenitätsgründen mußten die entsprechenden Daten aus den cechischen Listen ausgeklammert und auch in die Gruppe 'Sonstige' eingereiht werden.

Die Angaben über die Präpositionen im Finnischen schließen auch die der Postpositionen ein. Die Postpositionen treten im Finnischen viel häufiger auf als die Präpositionen.³⁾

Tab. 1: Wortklassenverteilung der Textwörter der Tschechischen und finnischen Sprache

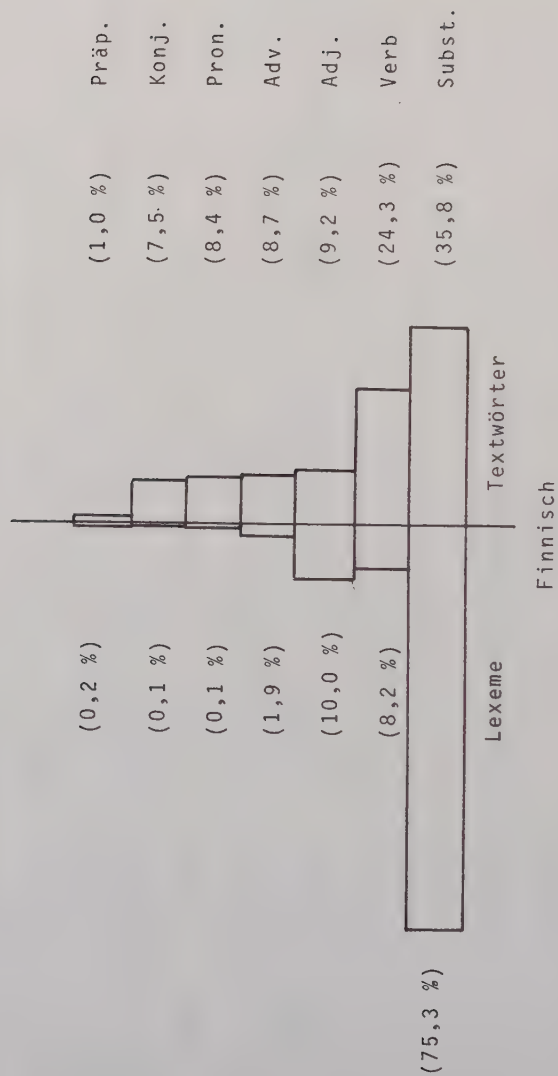
Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb	Präp. ^{x)}	Konj.	Sonstige
Finnisch in %	35,8	9,2	8,4	24,3	8,7	1,0	7,5	5,1
Tschechisch in %	27,8	11,2	11,0	18,2	10,3	10,1	9,8	1,7

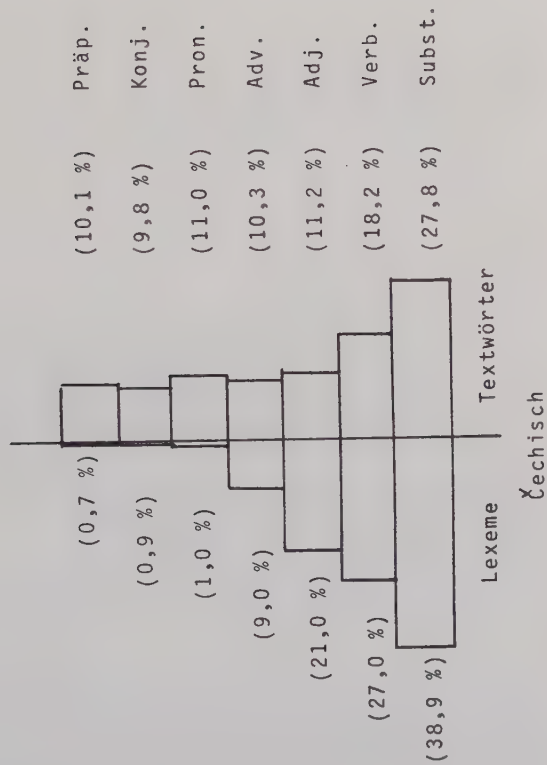
x) Im Finnischen einschließlich Postpositionen

Tab. 2: Wortklassenverteilung der Lexeme der tschechischen und finnischen Sprache

Wortklasse	Subst.	Adj.	Pron.	Verb	Adverb	Präp. x)	Konj.	Sonstige
Finnisch in %	75,3	10,0	0,1	8,2	1,9	0,2	0,1	4,2
tschechisch in %	38,9	21,0	1,0	27,0	9,0	0,7	0,9	1,5

x) Im Finnischen einschließlich Postpositionen





2. STILKATEGORIEN UND WORTKLASSENDISTRIBUTION

Jedes Wort gehört einer Wortklasse an. Dies ist schon ein Beleg für die Aufgabe dieses Wortes, als Satzteil zu fungieren. Damit wird eine Verteilung in die Wortklassen zu einem stilistischen Zeichen, das auf die Basis der lexikalischen Bedeutung verweist. Diese Verteilung wird immer vollzogen, aber nicht ausdrücklich reflektiert, stellt dennoch einen wichtigen stilistischen Bestandteil dar.⁴⁾

Die Diagramme (1 - 7) auf den Seiten 79 - 85 zeigen jeweils die Frequenzen einer Wortklasse innerhalb der vier Stil Kategorien, zusätzlich verglichen mit der Frequenz derselben Wortklasse im Gesamtmaterial des Finnischen und Čechischen. Die Čechische Kurve wurde mit einer unterbrochenen Linie, die finnische mit einer durchgezogenen Linie gezeichnet.

Die durch die Diagramme vermittelte Information bedarf einiger Erläuterungen:

Diagramm 1: Substantive

Sowohl in der Čechischen und in der finnischen Literatursprache (B) als auch in den gesprochenen Sprachen (E) treten prozentual weniger Substantive auf als im jeweiligen Gesamtmaterial, während das umgekehrte Verhältnis in den Zeitungssprachen (D) und Wissenschaftssprachen (C) vorliegt.

Diagramm 2: Adjektive

Das Verhältnis der Adjektive ist analog zu dem der Substantive in allen Stil Kategorien beider Sprachen.

Sowohl Herdan⁵⁾ als auch Mistrík hoben in ihren Untersuchungen dieses Ergebnis hervor: Je höher der Prozentsatz der Substantive ist, desto größer ist auch die Anzahl der Adjektive.

Diagramm 3: Pronomina

In der gesprochenen Sprache (E) und der Literatursprache (B) kommen sehr häufig Pronomina vor, in der Wissenschaftssprache (C) und der Zeitungssprache (D) dagegen nur selten. Dies trifft in beiden Sprachen zu.

Diagramm 4: Verben

Die geringste Anzahl der Verben ist in der Zeitungssprache (D) und Wissenschaftssprache (C), die größte in der Literatursprache (B) zu finden. Hier laufen die Kurven des Čechischen und des Finnischen in Bezug auf die Wissenschaftssprache (C) und die Zeitungssprache (D) auseinander. Der sonst parallele Verlauf beider Kurven ist in diesem Zusammenhang am deutlichsten gestört.

Diagramm 5: Adverbien

Die Adverbien sind in den beiden Sprachen und in allen Stil kategorien mehr oder weniger gleichmäßig verteilt.

Diagramm 6: Präpositionen

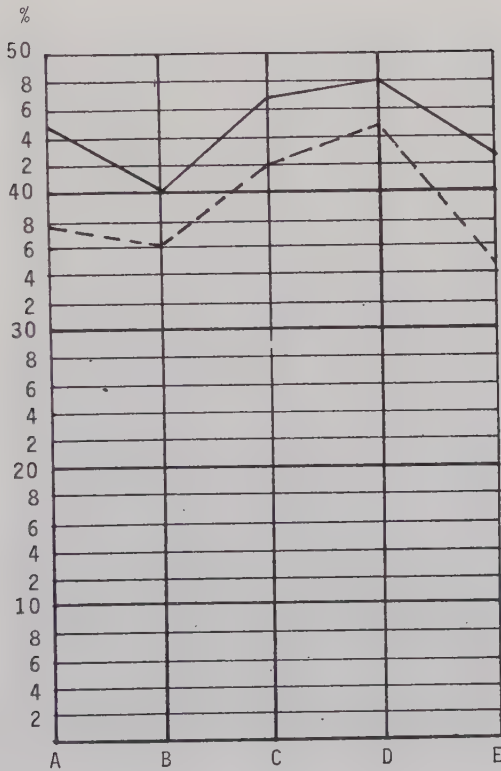
Durch alle Stil kategorien in beiden Sprachen treten die Präpositionen prozentual gleichmäßig verteilt auf, mit Ausnahme der Čechischen Zeitungssprache (D), bei der die Kurve deutlich steigt. Das geringe Auf-

treten der Präpositionen im Finnischen verursacht deren niedrige Prozentzahl (ca. 1 %).

Diagramm 7: Konjunktionen

Die Konjunktionen finden ihren niedrigsten Prozentsatz in der Zeitungssprache (D) beider Sprachen, aber den größten in der gesprochenen tschechischen Sprache (E) sowie in der Wissenschaftssprache (C) des Tschechischen, während ihr Prozentsatz in diesen Stil Kategorien sich im Finnischen nicht vom Durchschnitt abhebt.

Diagramm 1: Substantiva



— Finnisch

- - - Čechisch

A - Gesamtmaterial

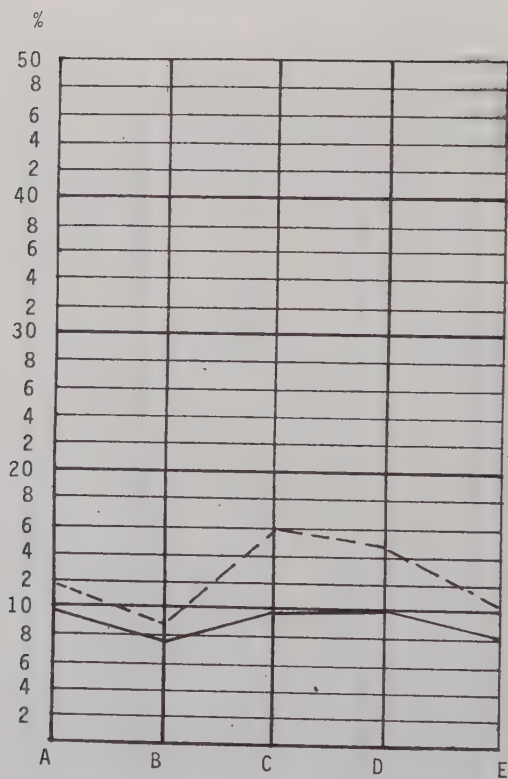
B - Literatursprache

C - Wissenschaftssprache

D - Zeitungssprache

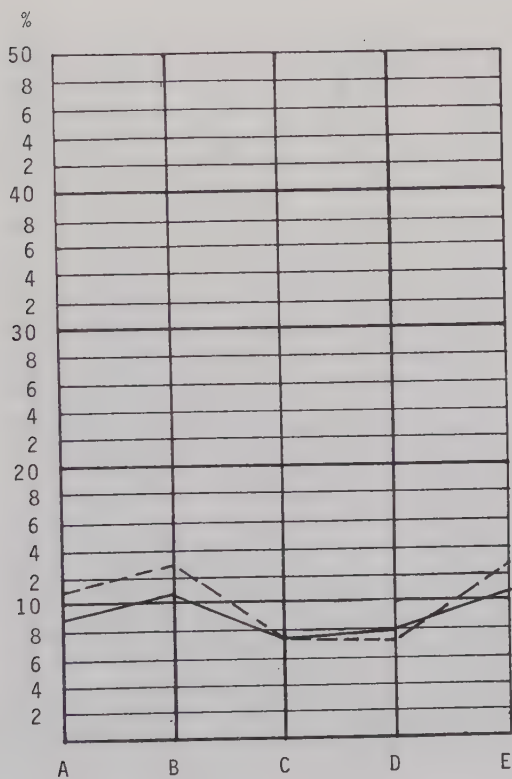
E - Gesprochene Sprache

Diagramm 2: Adjektiva



- Finnisch
--- Čechisch
- A - Gesamtmaterial
B - Literatursprache
C - Wissenschaftssprache
D - Zeitungssprache
E - Gesprochene Sprache

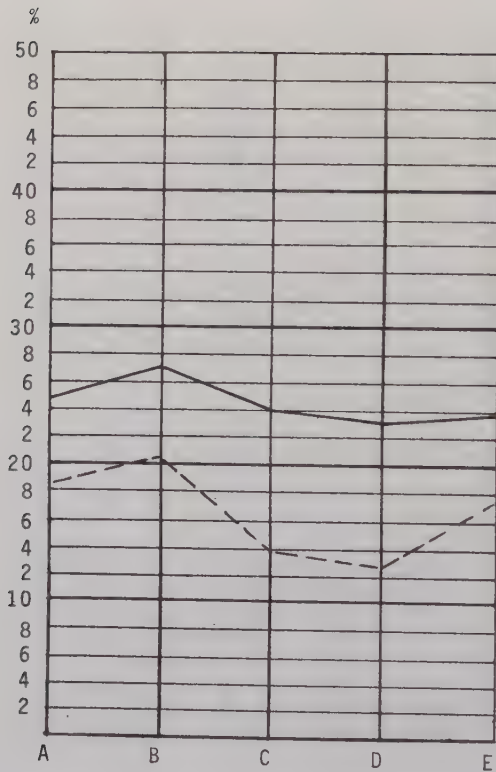
Diagramm 3: Pronomina



—— Finnisch
----- Čechisch

- A - Gesamtmaterial
- B - Literatursprache
- C - Wissenschaftssprache
- D - Zeitungssprache
- E - Gesprochene Sprache

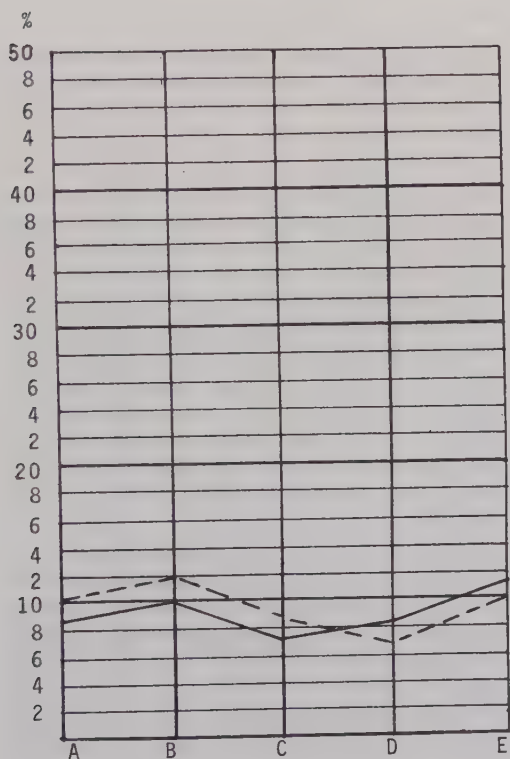
Diagramm 4: Verben



— Finnisch
- - - Čechisch

- A - Gesamtmaterial
- B - Literatursprache
- C - Wissenschaftssprache
- D - Zeitungssprache
- E - Gesprochene Sprache

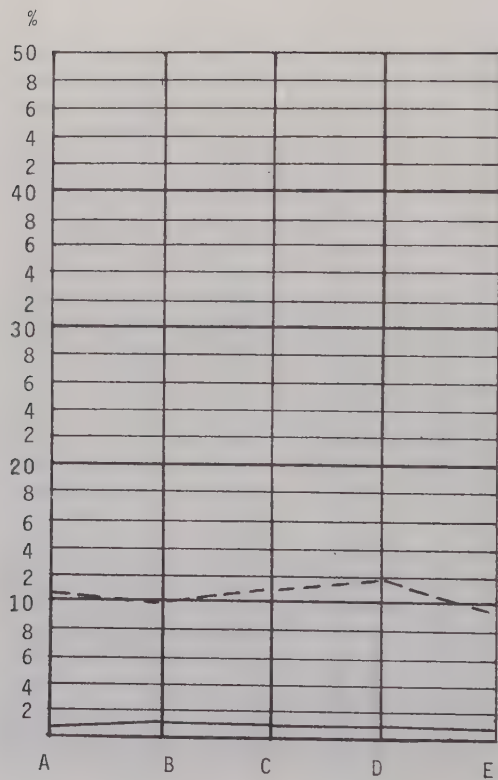
Diagramm 5: Adverbien



—— Finnisch
- - - - Čechisch

- A - Gesamtmaterial
- B - Literatursprache
- C - Wissenschaftssprache
- D - Zeitungssprache
- E - Gesprochene Sprache

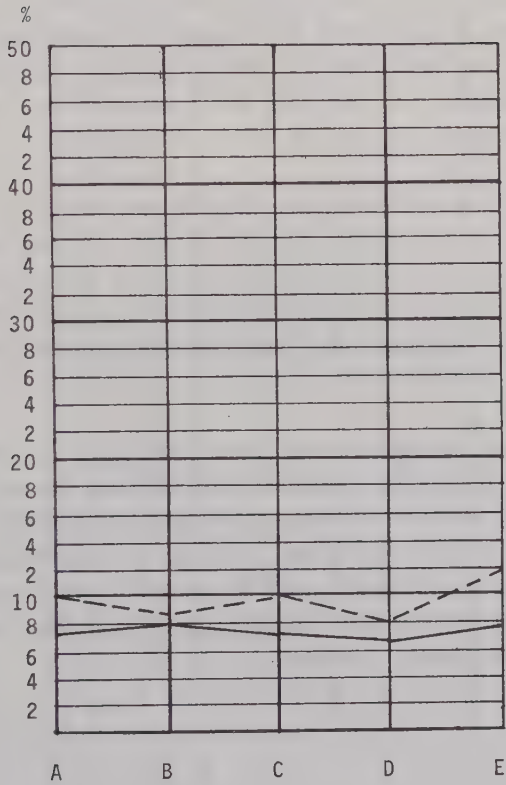
Diagramm 6: Präpositionen



— Finnisch
- - - Čechisch

- A - Gesamtmaterial
- B - Literatursprache
- C - Wissenschaftssprache
- D - Zeitungssprache
- E - Gesprochene Sprache

Diagramm 7: Konjunktionen



— Finnisch

- - - Čechisch

A - Gesamtmaterial

B - Literatursprache

C - Wissenschaftssprache

D - Zeitungssprache

E - Gesprochene Sprache

2.2. Textuale und lexikale Wortklassendistribution

Die textuale und lexikale Wortklassenverteilung in den verschiedenen Stil kategorien beider Sprachen hat eines gemeinsam: Substantive dominieren anteilmäßig sowohl in der Lexik als auch unter den Textwörtern. Nach Substantiven benutzt die Tschechische Zeitungs- und Wissenschaftssprache Adjektive am häufigsten, während das lexikalische Inventar der gesprochenen und der Literatursprache die Verben vor den Adjektiven bevorzugt.

Die Lexik der finnischen Literatursprache beinhaltet doppelt so viele Verben wie die Wissenschaftssprache, in der von allen Stil kategorien auch prozentual die meisten Substantive vorkommen. Auffallend ist überhaupt die Dominanz der Substantive im finnischen lexikalischen Inventar durch alle Stil kategorien.

Unter den Textwörtern dominieren ebenfalls die Substantive. Die Abweichungen der übrigen Wortklassen in Bezug auf die Stil kategorie sind unauffällig. Die Graphiken auf den Seiten 95 - 98 veranschaulichen deutlich die Proportionen der jeweiligen Wortklassen im lexikalischen Inventar wie im Rahmen der Textwörter in beiden Sprachen.

Tab. 3: Wortklassenverteilung der Textwörter in der cechischen und finnischen
Literatursprache

Wortklasse	Subst.	Adj.	Pron.	Verb	Adverb	Präp. x)	Konj.	Sonstige
Finnisch in %	30,0	7,8	11,2	27,6	10,3	1,2	8,0	3,9
Cechisch in %	26,2	8,4	13,0	20,5	11,0	9,9	9,1	1,9

x) Im Finnischen einschließlich Postpositionen

Tab. 4: Wortklassenverteilung der Textwörter in der Tschechischen und finnischen Wissenschaftssprache

Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. ^{x)}	Konj.	Sonstige
Finnisch in %	37,4	9,7	7,3	24,2	7,4	1,1	7,7	5,2
Tschechisch in %	32,1	16,1	7,3	14,0	9,0	10,5	9,7	1,3

x) Im Finnischen einschließlich Postpositionen

Tab. 5: Wortklassenverteilung der Textwörter in der Tschechischen und finnischen gesprochenen Sprache

Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. x)	Konj.	Sonstige
Finnisch in %	33,1	8,1	10,4	23,7	11,1	0,9	7,4	5,3
Čechisch in %	25,2	12,5	12,4	16,7	10,0	9,5	11,7	2,1

x) Im Finnischen einschließlich Postpositionen

Tab. 6: Wortklassenverteilung der Textwörter in der tschechischen und finnischen Zeitungssprache

Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. x)	Konj.	Sonstige
Finnisch in %	38,0	9,9	7,6	23,6	8,2	1,1	7,0	6,4
tschechisch in %	35,1	15,4	7,2	13,4	7,3	11,7	8,1	1,8

x) Im Finnischen einschließlich Postpositionen

Tab. 7: Wortklassenverteilung der Lexeme der Tschechischen und finnischen
Literatursprache

Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. x)	Konj.	Sonstige
Finnisch in %	58,3	11,8	0,4	19,7	5,3	0,6	0,2	3,7
Čechisch in %	38,2	17,7	1,3	30,3	9,1	0,9	1,0	1,5

x) Im Finnischen einschließlich Postpositionen

Tab. 8: Wortklassenverteilung der Lexeme der Tschechischen und finnischen Wissenschaftssprache

Wortklasse	Subst.	Adj.	Pron.	Verb	Ädverb	Präp. x)	Konj.	Sonstige
Finnisch in %	73,4	11,1	0,2	8,7	2,2	0,2	0,1	4,2
Čechisch in %	38,6	26,6	1,0	21,7	9,2	0,9	1,2	0,8

x) Im Finnischen einschließlich Postpositionen

Tab. 9: Wortklassenverteilung der Lexeme der Tschechischen und finnischen
Zeitungssprache

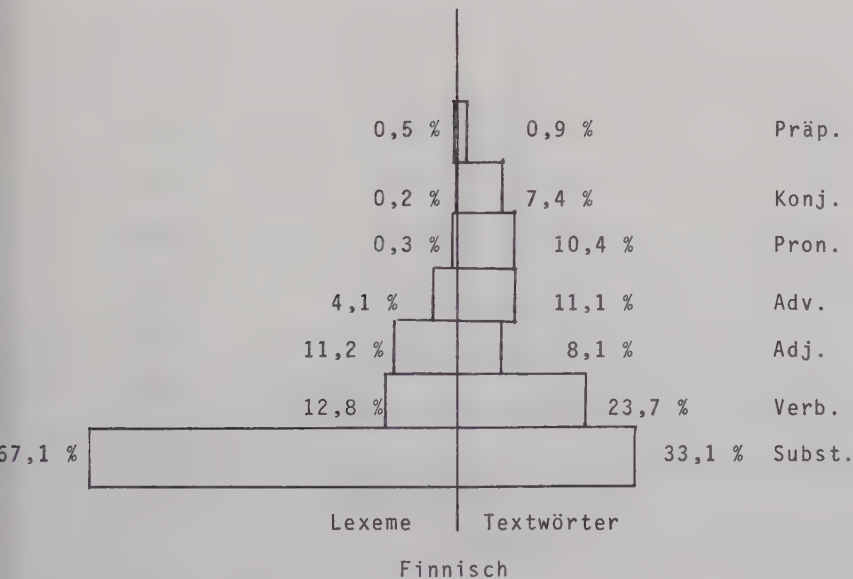
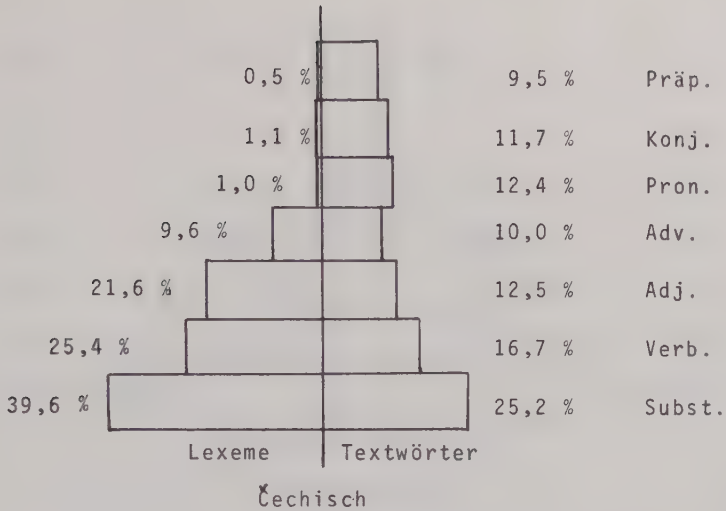
Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. x)	Konj.	Sonstige
Finnisch in %	70,5	11,2	0,2	10,7	3,0	0,3	0,1	4,0
Čechisch in %	47,3	23,5	0,9	18,0	7,7	0,7	0,9	1,0

x) Im Finnischen einschließlich Postpositionen

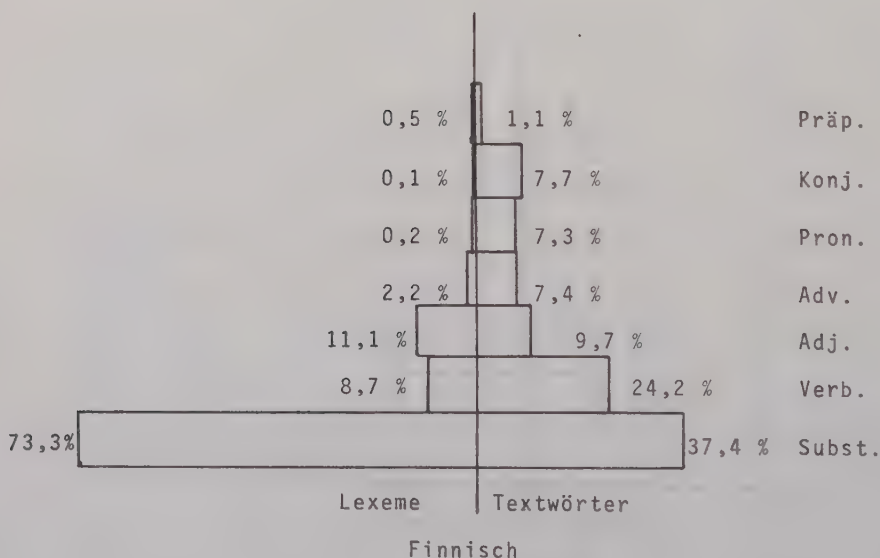
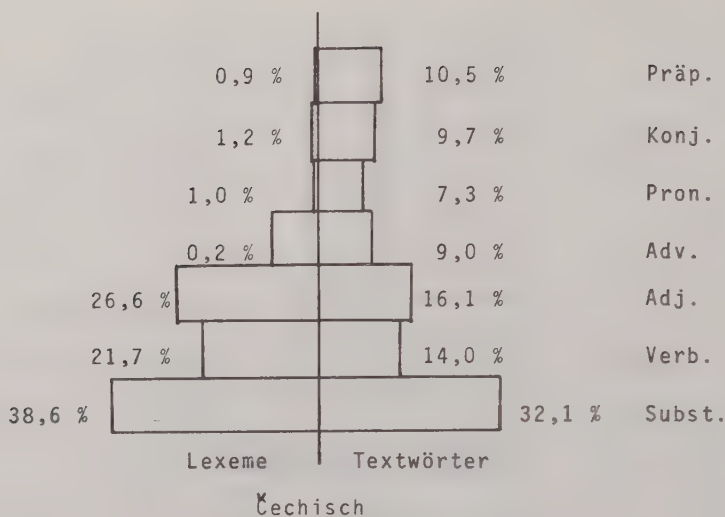
Tab. 10: Wortklassenverteilung der Lexeme der tschechischen und finnischen gesprochenen Sprache

Wortklasse	Subst.	Adj.	Pron.	Verb.	Adverb.	Präp. x)	Konj.	Sonstige
Finnisch in %	67,1	11,2	0,3	12,8	4,1	0,5	0,2	3,8
Čechisch in %	39,6	21,6	1,0	25,4	9,6	0,8	1,1	0,9

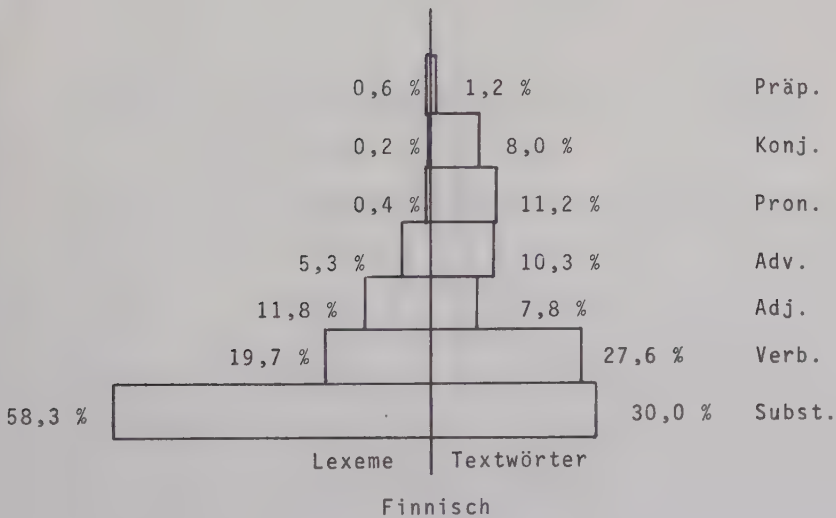
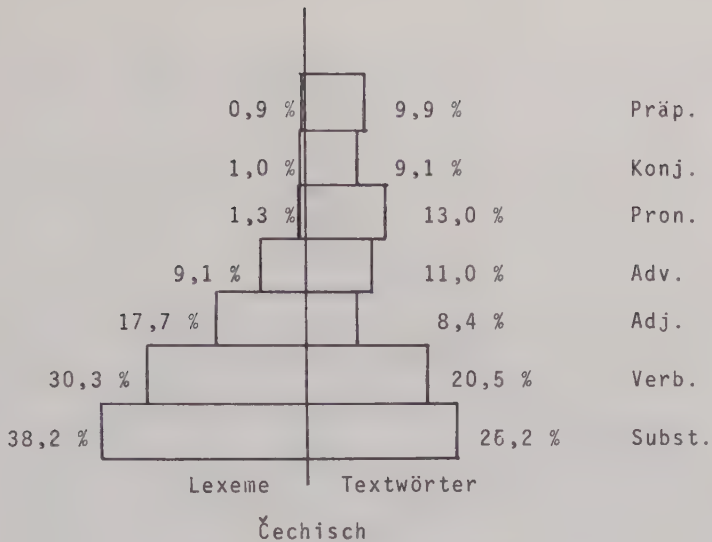
x) Im Finnischen einschließlich Postpositionen



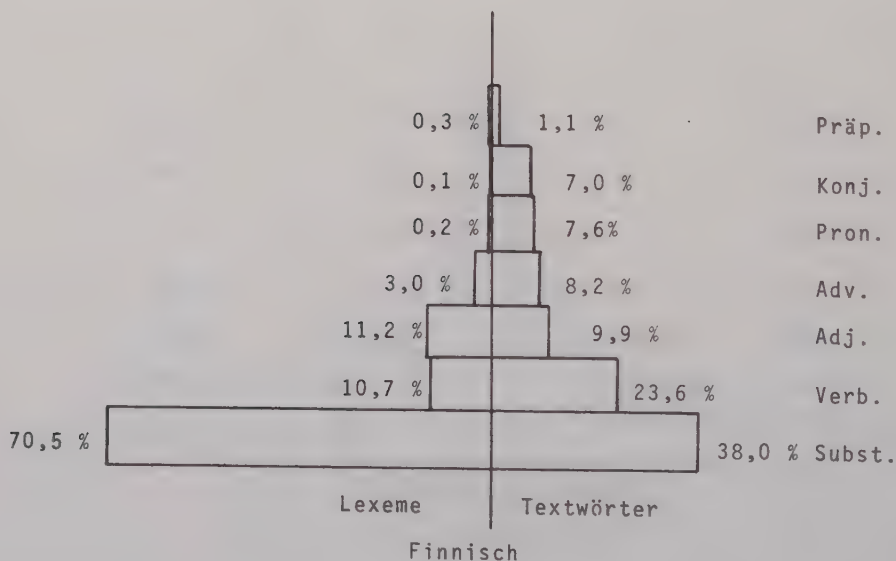
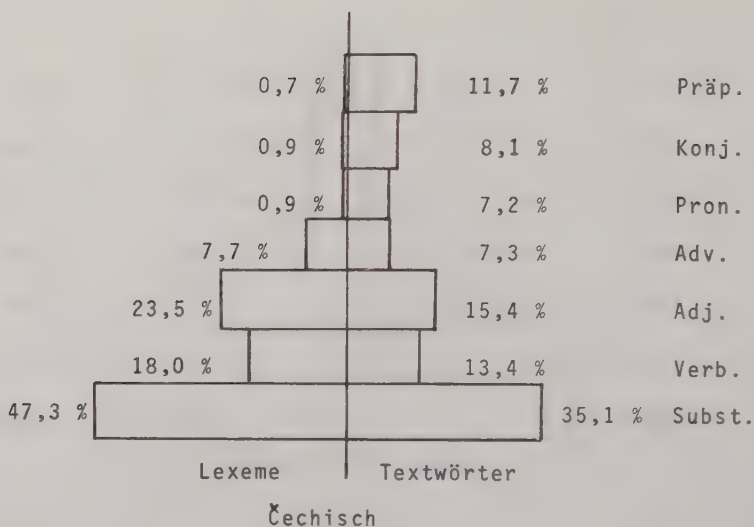
Wortklassenverteilung der Lexeme und Textwörter
in der gesprochenen Sprache



Wortklassenverteilung der Lexeme und Textwörter in der Wissenschaftssprache



Wortklassenverteilung der Lexeme und Textwörter
in der Literatursprache



Wortklassenverteilung der Lexeme und Textwörter
in der Zeitungssprache

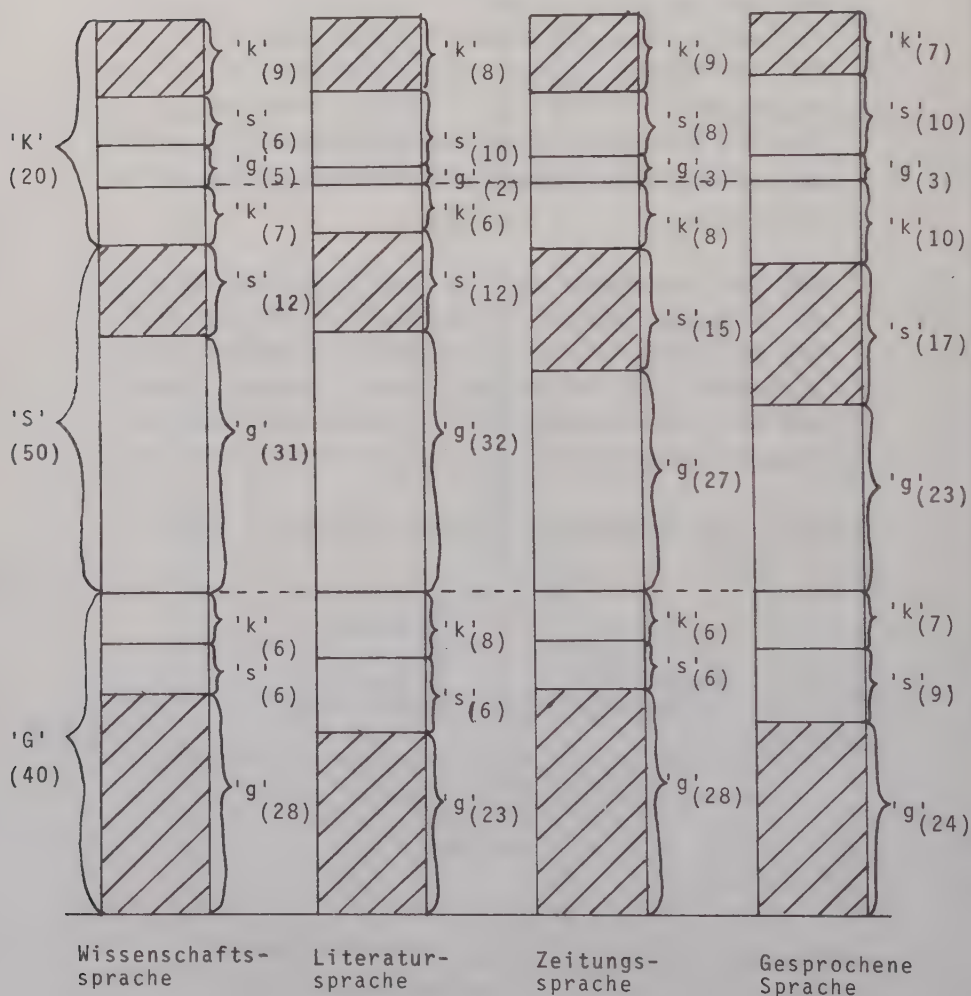
2.2. Stilkategorien und vertikale Wortdistribution

Die vertikale Wortdistribution gehört zwar nicht zu der grammatischen Statistik, sondern zu dem Bereich der Lexikostatistik, trotzdem soll hier der Vergleich der jeweils 110 häufigsten Wörter aus vier Stil kategorien durchgeführt werden. Als Beispiel werden die parametrischen Mengen der Literatur-, Zeitungs- und Wissenschaftssprache sowie der gesprochenen Sprache des Finnischen, in die diese Wörter - als Einteilungskriterium zuerst nach der Frequenz, dann nach der Semantik - eingeteilt werden, miteinander verglichen, um mögliche Abweichungen oder Übereinstimmungen festzustellen.

Aus der graphischen und der numerischen Darstellung der folgenden Seiten 100 und 101 kann man ersehen, wie in allen stilistischen Gruppen 'k'-Wörter normal vertreten sind, ebenso wie in der Literatursprache die verringerte Anzahl der 's'-Wörter und das verstärkte Vorkommen der 'g'-Wörter festzustellen ist.

(Vgl. lexikostatistische Vergleiche in dieser Arbeit, S. 58 ff)

Die graphische Darstellung der Frequenz-Semantik-Äquivalenz der Mengen 'K', 'S', 'G' und 'k', 's', 'g' in der finnischen Wissenschaftssprache, gesprochenen Sprache sowie in der Literatur- und Zeitungssprache:



Die numerische Darstellung der Frequenz-Semantik-Äquivalenz in der finnischen Literatur-, Zeitungs- und Wissenschaftssprache sowie in der gesprochenen Sprache:

Gliederung nach Frequenz	Menge 'K'			Menge 'S'			Menge 'G'		
Semantische Gliederung	'k'	's'	'g'	'k'	's'	'g'	'k'	's'	'g'
Wissenschafts- sprache	9	6	5	7	12	31	6	6	28
Literatur- sprache	8	10	2	6	12	32	8	9	23
Zeitungs- sprache	9	8	3	8	15	27	6	6	28
Gesprochene Sprache	7	10	3	10	17	23	7	9	24

V. TEXTDECKUNG

1. TEXTDECKUNG DER HÄUFIGSTEN WÖRTER

Das häufigste tschechische Wort 'a' kommt in dem für das Frequenzwörterbuch zugrunde gelegten Korpus von 1 623 527 Wörtern 67 122 mal vor, das zweithäufigste 'býti' 43 148 mal etc. Die Zahlen in der folgenden Liste entsprechen jeweils dem Vorkommen der zehn häufigsten Wörter im Gesamtkorpus, einmal einzeln, zum anderen kumulativ.

Wort	einzeln	kumulativ
a	67 122	67 122
býti	43 148	110 270
ten	37 280	147 550
v(e)	33 679	181 229
on	32 496	213 725
na	27 753	241 478
ze	18 092	259 570
s(e)	14 951	274 521
z(e)	13 408	287 929
který	11 692	299 621
Summe	299 621	

Verteilung der 100 häufigsten Wörter:

	einzeln	kumulativ
1. - 10. Wort	299 621	299 621
11. - 20. Wort	98 908	398 529
21. - 30. Wort	67 573	466 102
31. - 40. Wort	48 322	514 424
41. - 50. Wort	35 703	550 127
51. - 60. Wort	29 122	579 249
61. - 70. Wort	26 510	605 759
71. - 80. Wort	23 620	629 379
81. - 90. Wort	17 683	649 326
91. - 100. Wort	17 683	667 009

Das Verhältnis dieser Zahlen zum Gesamtkorpus ergibt die prozentualen Einzel- und Kumulativangaben.

Demnach decken die 10 häufigsten Wörter im Čechischen 18,5 %, im Finnischen 17,5 % eines beliebigen Textes, die weiteren zehn zusammen repräsentieren im Čechischen 6,1 % und im Finnischen 4,2 %, die weiteren zehn im Čechischen 4,2 %, 2,9 % im Finnischen, und diese 30 häufigsten decken 28,8 % eines Čechischen und 24,6 % eines finnischen Textes. Die 100 häufigsten dagegen sind mit 41,2 % der Wörter im Čechischen und mit 35,1 % im Finnischen deckungsgleich.

Auffallend ist die Gleichmäßigkeit der prozentualen Verkleinerung der Textdeckung in den beiden Sprachen, sowohl bei Intervallen von je 100 als auch bei Intervallen von je 10.

Erweitert man die Zahl der häufigsten Wörter auf 1000, so steigt die Textdeckung im Čechischen auf 66,1 % und im Finnischen auf 64,8 %.

Auf der Seite 104 wird eine prozentuale Übersicht der Textdeckung von den 100 häufigsten Čechischen und finnischen Wörtern in Intervallen von zehn Wörtern aufgestellt. Auf der Seite 105 ist das entsprechende von den 1 000 häufigsten Čechischen Wörtern in Intervallen von 100 Wörtern zu entnehmen, und auf der Seite 106 stehen die gleichen Angaben über die finnische Sprache.

Prozentuale Übersicht der Textdeckung von den 100 häufigsten Čechischen und finnischen Wörtern, einzeln und in Intervallen von je zehn

Wort	Einzel in %		Kumulativ in %	
	Čechisch	Finnisch	Čechisch	Finnisch
1. - 10. Wort	18,5	17,5	18,5	17,5
11. - 20. Wort	6,1	4,2	24,6	21,7
21. - 30. Wort	4,2	2,9	28,8	24,6
31. - 40. Wort	3,0	2,3	31,8	26,9
41. - 50. Wort	2,2	1,9	34,0	28,8
51. - 60. Wort	1,8	1,5	35,8	30,3
61. - 70. Wort	1,6	1,3	37,4	31,6
71. - 80. Wort	1,5	1,1	38,9	32,9
81. - 90. Wort	1,2	1,1	40,1	34,0
91. - 100. Wort	1,1	1,1	41,2	35,1

Übersicht über die prozentuale Textdeckung der 1 000 häufigsten Wörter des
Čechischen, einzeln und in Intervallen von je hundert

Wörter	die ersten 100	die zweiten 100	die dritten 100	die vierten 100	die fünft en 100	die sechsten 100	die siebten 100	die achten 100	die neunten 100	die zehnten 100
Einzeln in %	41,2	7,0	4,2	3,1	2,5	2,0	1,9	1,4	1,4	1,2
Kumulativ in %	41,2	48,2	52,4	55,5	58,0	60,0	61,9	63,5	64,9	66,1

Übersicht über die prozentuale Textdeckung der 1 000 häufigsten Wörter des Finnischen, einzeln und in Intervallen von je hundert

Wörter	die ersten 100	die zweiten 100	die dritten 100	die vierten 100	die fünften 100	die sechsten 100	die siebten 100	die achten 100	die neunten 100	die zehnte 100
Einzeln in %	35,1	7,8	5,1	3,9	3,1	2,6	2,2	1,7	1,9	1,4
Kumulativ in %	35,1	42,9	48,0	51,9	55,0	57,6	59,8	61,5	63,4	64,8

Übersicht über die prozentuale Textdeckung der 1 000 häufigsten Wörter des
Čechischen, einzeln und in Intervallen von je hundert

Wörter	die ersten 100	die zweiten 100	die dritten 100	die vierten 100	die fünften 100	die sechsten 100	die siebten 100	die achten 100	die neunten 100	die zehnten 100
Einzeln in %	41,2	7,0	4,2	3,1	2,5	2,0	1,9	1,4	1,4	1,2
Kumulativ in %	41,2	48,2	52,4	55,5	58,0	60,0	61,9	63,5	64,9	66,1

Übersicht über die prozentuale Textdeckung der 1 000 häufigsten Wörter des Finnischen, einzeln und in Intervallen von je hundert

Wörter	die ersten 100	die zweiten 100	die dritten 100	die vierten 100	die fünften 100	die sechsten 100	die siebten 100	die achten 100	die neunten 100	die zehnte 100
Einzel in %	35,1	7,8	5,1	3,9	3,1	2,6	2,2	1,7	1,9	1,4
Kumulativ in %	35,1	42,9	48,0	51,9	55,0	57,6	59,8	61,5	63,4	64,8

Die Textdeckung kann anschaulich durch eine interpoliert gezeichnete Kurve dargestellt werden.

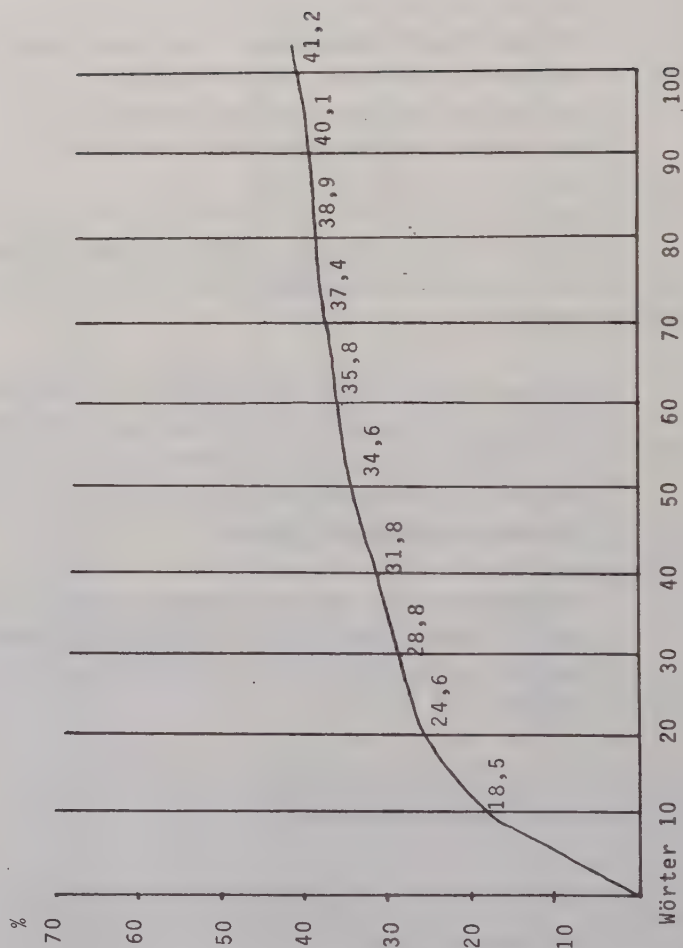
Auf der Seite 108 zeigt diese Kurve die prozentuale Textdeckung in Abhängigkeit von der Anzahl der 100 am häufigsten verwendeten Čechischen Wörter in Intervallen von je zehn, auf der Seite 109 befindet sich die entsprechende Kurve des Finnischen.

Der interpolierte Kurvenverlauf wird ab dem zwanzigsten Wort der Frequenzreihe in beiden Sprachen fast horizontal, danach folgt nur noch eine leichte Steigung.

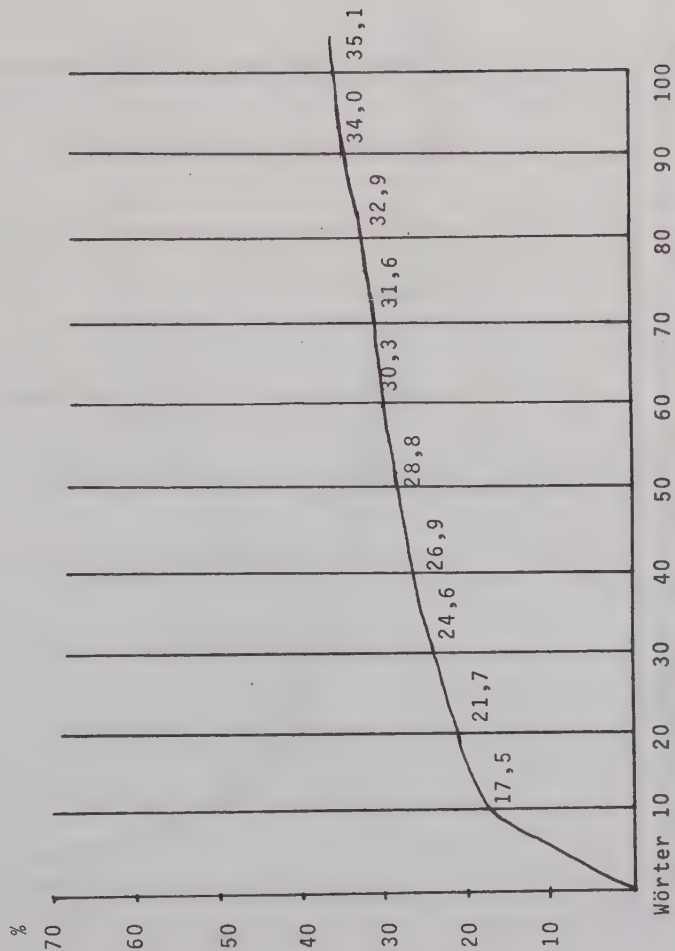
Auf den Seiten 110 + 111 zeigen die Kurven die prozentuale Textdeckung in Abhängigkeit von der Anzahl der 1 000 häufigsten Čechischen bzw. finnischen Wörter in Intervallen von je hundert. Diese Kurven verlaufen annähernd gleich.

Zusammenfassend kann man sagen, daß der Verlauf der Kurven approximativ ist. Unterschiede lassen sich am ehesten noch bei den 10-er Intervallen feststellen.

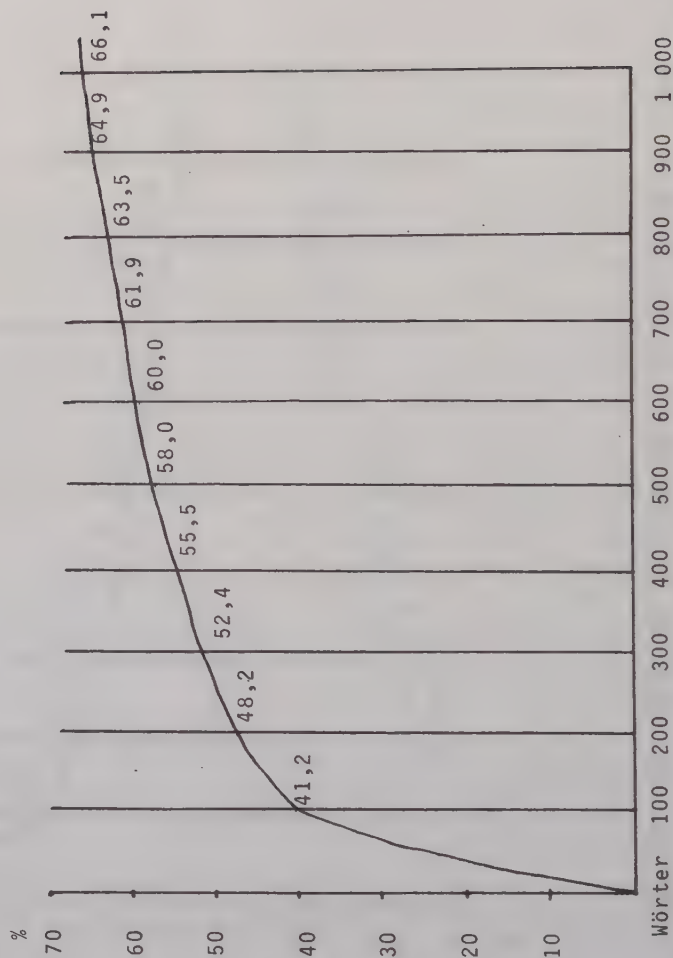
Abhängigkeit der Textdeckung in % von der Anzahl der 100 häufigsten tschechischen Wörter in Intervallen von je 10, interpolierter Kurvenverlauf



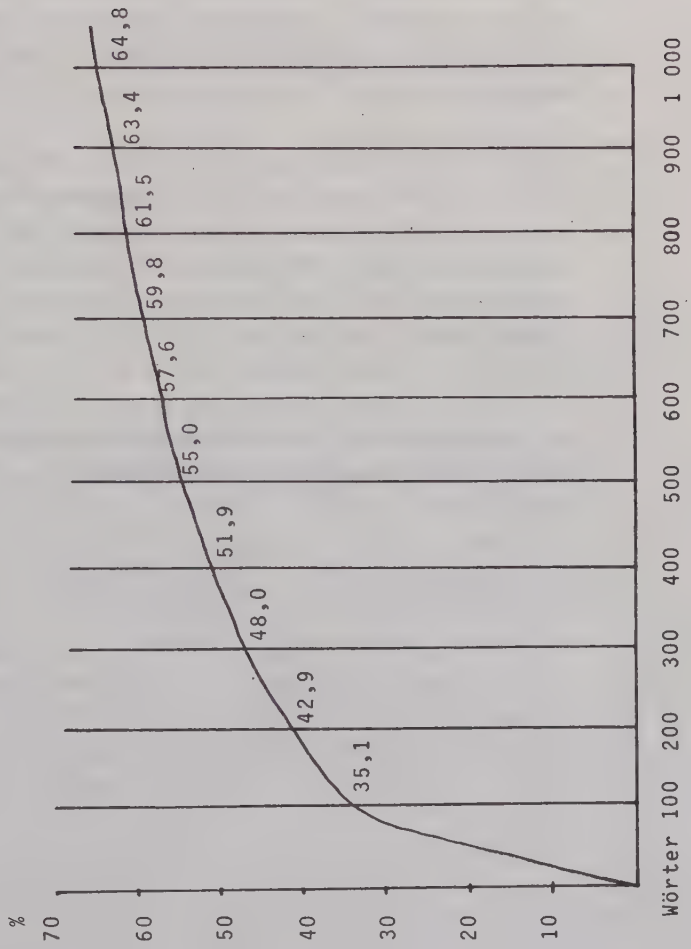
Abhängigkeit der Textdeckung in % von der Anzahl der 100 häufigsten finnischen Wörter in Intervallen von je 10, interpolierter Kurvenverlauf



Abhängigkeit der Textdeckung in % von der Anzahl der 1 000 häufigsten tschechischen Wörter in Intervallen von je 100, interpolierter Kurvenverlauf



Abhängigkeit der Textdeckung in % von der Anzahl der 1 000 häufigsten finnischen Wörter in Intervallen von je 100, interpolierter Kurvenverlauf



2. TEXTDECKUNG IM LAUFENDEN TEXT

Daß im Čechischen die 100 häufigsten Wörter tatsächlich 40 % und mehr, die 1 000 häufigsten Wörter 65 % und mehr eines beliebigen Textes decken, soll hier an einem stichprobenartig ausgesuchten Textabschnitt überprüft werden.¹⁾

Die Wörter, die sich nicht innerhalb der 1 000 häufigsten befinden, sind mit der durchgezogenen Linie unterstrichen und jene Wörter, die den Wörtern mit dem Rang 101 bis 1 000 angehören, sind mit einer unterbrochenen Linie versehen und die zu den 100 häufigsten gehörenden Wörter sind ohne irgendeine Markierung.

Der Textabschnitt ist ein Auszug aus dem Roman 'Český román' von Olga Scheinpflugová.²⁾

"V Karlovyých Varech se usadili v tichém hotelu v postranní ulici; dva pokoje vedle sebe, dva rukopisy na psacích stolech. Druh slyšel druhá, jak přechází sem a tam, když nabírá nový rozběh myšlenek. Nikdy si nebyli tak blízko ve svých zvycích i ve svém soustředění, jako za těmito společnými zavřenými dveřmi. Po práci chodili celě hodiny po cestách a hladkých pásech silnic a mluvili o všem starém nějakými novými hlasy. Role, které byly odehrány, knihy už napsané byly jen lávkami, po nichž se přejde na nový břeh. Všecko, čeho se dotkli společným zájmem, bylo poznámeno společenstvím lásky a dohody, umění, svět i život bylo viděno z nějaké jiné strany nebo ze dvou najednou, často se okouzleně rozezmáli a nadchli, jako by byli objevili cosi docela nového. Věci starých strachů ztékali odvahou, dávných sporů se dotýkali s plachostí. Stalo se cosi velikého, co změnilo tvář jejich světa a skutečnosti, narodili se pro sebe před týdnem, nebot' patřice si naprosto patnáct roků, dohodli se nyní stejně jako před nimi, že se vezmou. Byly to nové, slavné námluvy. Ona si všimla sebe same s jakousi št'astnou ironií, jak se malebněji chová při chůzi, jindy tak klackovitě po jehu boku, jak pro něho osvěžila svoje unavené názory."³⁾

Aus dem am obigen Textabschnitt durchgeführten Experiment, d.h. an einem Korpus von 185 Wörtern, geht als Resultat hervor, daß die 100 häufigsten Wörter 45 %, die 1 000 häufigsten sogar 75 % dieses Textes decken. Weiter geht daraus hervor, daß 23 % der Wörter die Frequenz 196 oder darunter haben, was gleichzeitig bedeutet, daß diese Wörter den Rang 1 001 und höher in dem tschechischen Frequenzwörterbuch besetzen.

Aus den vorangegangenen Berechnungen geht hervor, daß die Textdeckung durch die 100 häufigsten Wörter des Finnischen wiederum 35 % und die durch die 1 000 häufigsten Wörter 65 % sein sollte.

Die Bestätigung dafür liefert das am Beispiel eines Textabschnittes von E.Joenpelto⁴⁾ durchgeführte Experiment, wobei 53 % der Wörter dieses Textes unter den 100 häufigsten und 74 % der Wörter unter den 1 000 häufigsten Wörtern zu finden sind.

Die Markierung ist dieselbe wie beim tschechischen Textbeispiel:

"Joku ihmetteli, mitä sellaisen tuli tehdä, mutta kappalainen arveli, että sen näyttäisi aika. Piti vain nimetä ihmisiä, mieluummin sellaisia, joilla ei ollut kovin paljon luottamustoimia tähän mennessä. Sillä niin oli, että liian monet joutuivat liian paljon uhraamaan aikaansa kunnan hommiin saamatta minkäänlaista, ei ainakaan edes kohtuullista korvausta uutteruudesta. Kappalainen kirjoitti nimiä sen mukaan kuin niitä keksittiin. Hänen tuli tuskainen tai pirullinen olo, kun hän ajatteli ettei siinä asiassa tunnuettu ollenkaan välitettävän maineesta. Hän oli pyytää puheenvuoroa, mutta diakonissa puhui jo hänen puolestaan. Vieno luimisteli ja katsoi pöytään. Ei hän mitään ollut tehnyt, hänelle jotkut tuollaiset orvot -. Rauhassa hän olisi halunnut olla. Tai tehdä jotakin muuta. Mitä helvettiä hän ymmärsi lapsenhoidosta? Eikö siinä pitänyt olla joku, jolla itsellä olisi lapsia. Tai ainakin olisi koulutettu lasten kanssa olemaan. Niin no, joku opettaja tai sairaanhoidtaja tai joku, kuka tahansa muu kuin yksinäinen miesihminen, joka-. Ja hänelle selvisi, että punakaartilaisen maine ei ehkä ollutkaan enää niin paha kuin esimerkiksi sellaisen naisen, jota vaikkapa epäiltiin uskotto- muudesta ja löysästä elämästä, oli muuten miten hyvä ihminen tahansa. Ihmisethän ovat pähkähulluja, hän ajatteli."⁵⁾

VI. ZUR SEMANTIK DER WÖRTER IN DER
FREQUENZLISTE

1. FREQUENZUNTERSCHIEDE UND POLYSEMIE

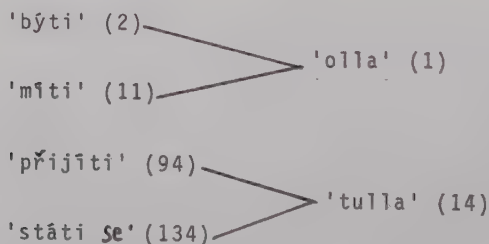
Ein wichtiger Faktor, der die Höhe der Frequenz eines Wortes beeinflusst, ist die Polysemie.

Das finnische Verb 'olla' beispielsweise, das in der finnischen Frequenzliste den Rang 1 besetzt, deckt die Bedeutung der tschechischen Verben 'mýti' und 'býti', deren die auf den Rängen 1 beziehungsweise 11 stehen.

Analog dazu wird dieses Problem noch deutlicher durch das Beispiel der Verben 'přijíti' und 'státi se' erläutert, denen nur ein Verb im Finnischen entspricht, und zwar 'tulla'. Das finnische 'tulla' besetzt den Rang 16, die tschechischen Äquivalenzen 94 bzw. 134.

Dasselbe läßt sich an der Frequenz der Berufsbezeichnungen zeigen, die im Finnischen für beide Geschlechter nur in der maskulinen Form gebraucht werden, obwohl die femininen Formen in diesem Falle existieren. Es gibt aber eine Reihe von Berufsbezeichnungen, die nur eine Form für beide Geschlechter hat.

Die feminine Form hat auch im tschechischen eine minimale Frequenz im Vergleich zur Frequenz der maskulinen Form.



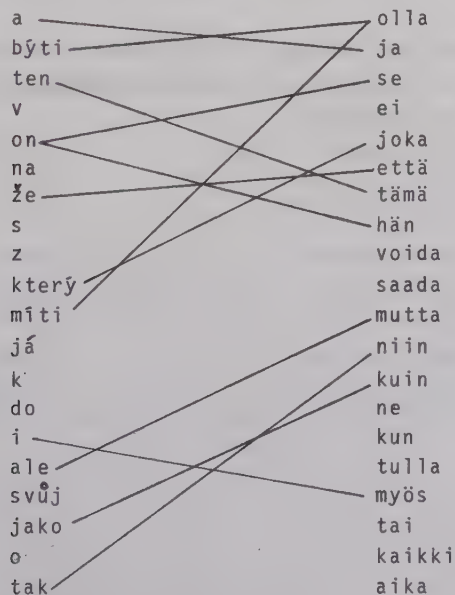
oder



Vergleicht man die semantischen Entsprechungen der ersten 20 häufigsten Wörter beider Sprachen miteinander, so sieht man, daß dabei 12 von ihnen bedeutungsmäßig identisch sind.

Wie schon die in dieser Arbeit durchgeführte vertikale Wortdistribution zeigte, gehören die hier semantisch abweichenden Wörter im Čechischen zu den Präpositionen, die im Finnischen so gut wie gar nicht gebraucht werden, im Finnischen kommen hier sogar einige autosemantische Wörter vor, die immer von der Auswahl mehr oder weniger abhängig sind.

In den Listen sind die Äquivalenzen mit durchgezogenen Linien miteinander verbunden.



Die semantische Inkongruenz beider Sprachen und deren Rückwirkung auf die Frequenz drückt sich z.B. auch bei der Segmentierung des Tages aus. Der Zeitabschnitt von 24 Stunden, Tag und Nacht - im Finnischen gibt es dafür eine eigene Bezeichnung 'vuorokausi' - gliedert sich im Čechischen und Finnischen, wie folgt:

yö (122)	aamuyö (3)	aamu (111)	
o	2	4	6 8
noc (925)		jítro (93)	
aamupäivä (20)	keskipäivä (8)	iltapäivä (26)	
10	12	14	16
dopoledne (23)	poledne (98)	odpoledne (64)	
ilta (137)	iltayö ¹⁾	keskiyö (5)	
18	20 22	24	
večer (433)		půlnoc (63)	

Hier wird die Frequenz stark durch Faktoren wie Lebensweise, Umwelt, geistige Orientierung des Sprachbenutzers beeinflusst.

2. ADJEKTIVE UND IHRE FREQUENZ

2.1. Adjektive mit Antonymen

Laut Mistrík zeigt sich im Slovakischen²⁾ bei der Frequenzbestimmung jener Adjektive, die Antonyme haben, deutlich die Tendenz, daß das Positive, das Ursprüngliche, das Größere des Oppositionspaares immer die höhere Frequenz aufweist.

Paralleles läßt sich auch in den hier untersuchten Sprachen feststellen.

Hier eine Tabelle mit Adjektiven, von denen eine Reihe doppelten Antonyme besitzt.

nový (2437) + mladý (787)	uusi (707 + nuori (401
starý (1647)	vanha (417)
veliký (2507)	suuri (1262) + iso (89)
malý (1081)	pieni (580)
dobrý (1174)	hyvä (630)
zlý (335) + špatný (203)	huono (77) + paha (130)
vysoký (506)	korkea (193) + syvä (87)
nízký (127)	matala (42)
těžký (501)	raskas (66)
lehký (180) + snadný (68)	kevyt (47)
mnoho (927)	paljon (505)
málo (429)	vähän (197)

2.2. Farbadjektive

Die Farben Schwarz, Weiß und Rot in ihrer Frequenz nach geordnet sind auch in anderen Sprachen³⁾ dominierend; die Reihenfolge innerhalb der dreien ist zwar unterschiedlich.

Die höhere Frequenz des tschechischen Adjektivs 'rudý' im Verhältnis zu der des Adjektivs 'červený' hängt sicher von Ausdrücken, wie 'Rudé právo', 'Rudá armáda', 'Rudé moře' etc. ab. Die Differenzierung in zwei verschiedene Bezeichnungen für die Farbe Rot im tschechischen schlägt sich in der Frequenz so nieder, daß die Frequenzsumme beider Bezeichnungen im Verhältnis der Frequenz des finnischen Äquivalents entspricht, die Reihenfolge in der Frequenzliste bleibt nach der Zusammenziehung trotzdem unverändert.

bílý (604)	musta (92)
černý (218)	punainen (80)
rudý (218)	valkoinen (78)
zelený (205)	vihreä (44)
červený (198)	sininen (31)
modrý (198)	keltainen (24)
žlutý (96)	ruskea (23)
hnědý (60)	

3. HIMMELSRICHTUNGEN, WOCHENTAGE, MONATS-
BEZEICHNUNGEN UND IHRE FREQUENZ

Listet man die Bezeichnungen der einzelnen Himmelsrichtungen jeweils nach ihrer Frequenz auf, verbindet die semantischen Entsprechungen mit Linien miteinander, so sieht man keine Parallelität in den Reihenfolgen:

západ (156)	—	pohjoinen (61)
východ (127)	—	länsi (36)
sever (102)	—	etelä (35)
jih (82)	—	itä (17)

Bei Wochentagen wiederum haben beide Sprachen die Bezeichnung für Sonntag am häufigsten und die für Samstag steht als an zweiter Stelle. Die übrigen Wochentage weisen keine Parallelitäten auf.

neděle (251)	—	sunnuntai (61)
sobota (82)	—	lauantai (31)
pátek (64)	—	maanantai (25)
středa (55)	—	tiistai (23)
pondělí (52)	—	perjantai (22)
čtvrtek (37)	—	keskiviikko (19)
úterý (36)	—	torstai (18)

Innerhalb der nach ihrer Frequenz aufgereihten Monatsbezeichnungen findet man keine Symmetrie. Die höchste Frequenz des Monats Juni im Finnischen kann möglicherweise mit der geographischen Lage Finnlands erklären, durch die dieser Monat besonders wichtig wird.

září (204)	kesäkuu (43)
říjen (78)	heinäkuu (37)
duben (69)	toukokuu (35)
květen (67)	elokuu (34)
únor (66)	lokakuu (34)
srpen (66)	joulukuu (34)
listopad (60)	tammikuu (30)
březen (58)	huhtikuu (30)
leden (50)	marraskuu (29)
červen (39)	helmikuu (25)
červenec (37)	syyskuu (25)
prosinec (31)	maaliskuu (20)

4. STAATSBEZEICHNUNGEN UND STÄDTENAMEN UND IHRE FREQUENZ

4.1. Staatsbezeichnungen mit der Frequenzangabe

Německo (204)	Suomi (659)
Československo (152)	Ruotsi (172)
Francie (134)	Neuvostoliitto (121)
Italie (116)	Tšekkoslovakia (77)
Amerika (106)	Saksa (72)
Polsko (41)	Englanti (64)

Deutlich sieht man, wie die Frequenz der Staatsbezeichnungen durch die geographische Nähe, die wirtschaftliche und kulturelle Bedeutung umliegender Staaten beeinflußt wird.

Verwunderlich ist zwar die hohe Frequenz des Wortes 'Tšekkoslovakia' im finnischen Frequenzwörterbuch. Allerdings, wenn die Zeitspanne des Exzerpierens (1967 - 1968) aus dem für das Wörterbuch zugrunde gelegten Material in Betracht gezogen wird, ist dies erklärbar, da das Wort durch die politischen Ereignisse Ende der 60er Jahre mindestens in einer von den vier Stil Kategorien, nämlich in der Zeitungssprache, häufiger als sonst auftauchte.

4.2. Städtenamen und ihre Frequenzen

Zwischen der Frequenz der Städtenamen und der Einwohnerzahl läßt sich auch eine Gesetzmäßigkeit andeutungsweise beobachten. Die Stadt mit der größten Einwohnerzahl hat auch die höchste Frequenz, die zweitgrößte Stadt oft die zweithöchste Frequenz etc.

Neben den Städtenamen sind die Frequenzangabe und die Einwohnerzahl aufgeführt:

Städtename	Frequenz	Einwohnerzahl
Praha	898	1 078 096 ⁴⁾
Brno	126	335 918
Kladno	32	153 408 ⁵⁾
Ostrava	21	278 737
Helsinki	265	490 205 ⁶⁾
Turku	94	165 004
Tampere	89	165 418
Oulu	44	93 218

5. NUMERALIA UND IHRE FREQUENZ

Die Kardinalzahlen von 1 bis 5 im Čechischen ordnen sich so, daß die Zahl 1 die höchste, die Zahl 2 die zweithöchste Frequenz aufweist und ähnlich bis zur Zahl 5. Im Finnischen hat die Zahl 2 eine höhere Frequenz als die Zahl 1 und die Zahl 5 eine höhere als die Zahl 4.

jeden (3082)	————	kaksi (592)
dva (2015)	————	yksi (522)
tři (801)	————	kolme (330)
čtyři (360)	————	viisi (141)
pět (297)	————	neljä (139)

Bei den Ordinalzahlen ist die Reihenfolge der semantischen Äquivalenzen beider Sprachen identisch: die Bezeichnung für das Zweite hat eine höhere Frequenz als die für das Erste.

druhý (1813)	————	toinen (1204)
první (1555)	————	ensimmäinen (535)
třetí (341)	————	kolmas (145)
čtvrtý (102)	————	neljäs (55)
pátý (94)	————	viides (27)

SCHLUßBEMERKUNG

Als wichtiges Ergebnis der in dieser Untersuchung durchgeführten lexikostatistischen Vergleiche zeigt sich, daß das Fehlen der Präpositionen im Finnischen eine Verschiebung der autosemantischen Wörter unter die häufigsten Wörter zufolge hat, während diese im Tschechischen unter weniger-häufiger-vorkommende Wörter fallen.

Bemerkenswert ist noch das durch grammatisch-statistische Vergleiche gewonnene Ergebnis, daß die Prozentsatzhöhe der dominierenden Wortklasse - Substantive - im Finnischen erheblich höher ist, als die des Tschechischen, was auch dadurch beeinflußt wird, daß ein Wort im Finnischen mehreren Wortklassen angehören kann.¹⁾

Abschließend muß noch einmal darauf hingewiesen werden, daß eine quantitative Information immer nur Signal eines inhaltlichen Phänomens ist.

"Die exakte Information über die Distribution eines Textelementes darf nicht zu einem komplexen Kriterium erhöht, sondern muß als Signal verstanden werden, das auf die potentielle Anwesenheit einer Textqualität hinweist."²⁾

ANMERKUNGEN

Die Daten für die Vergleiche wurden den folgenden Frequenzuntersuchungen entnommen:

1. Jelínek, Bečka, Těšitelová; Frekvence slov, slovních druhů a tvarů v českém jazyce, Praha 1961
2. Saukkonen, Haipus, Niemikorpi, Sulkala, Suomen kielen taajuussanasto, Porvoo 1979
3. Mistrík, Frekvencia slov v slovenčine, Bratislava 1969
4. Kaasik, Tuldava, Villup, Ääremaa, Eesti keele ilukirjandusproosa autorikõne sagedussõnastik 1 und 2. in: Keelestatistika 1, Tartu 1976, Keelestatistika 2, Tartu 1977

Hier wurden unter dem wissenschaftlichen Stil die Daten von 'odborná literatura' und 'vedecká literatura' zusammengerechnet. Auch die Daten für Literatursprache sind die Summe von 'beletrie', 'poesie', 'literatura pro mládež', 'dramata'.

Kapitel_I

1. Herdan, 1964, S.3
2. Muller, 1972
3. Cohen, Sur la statistique..., S. 8 - 16
4. Spang-Hansen, S. 133 ff.
5. Těšitelová, Otázky..., S. 9
6. ebd., S. 11
7. Dietrich, Klein, S. 115
8. Bense, S. 109 ff.

9. Bar-Hillel 1960, 1967
Josselson 1971 und Mel'čuk, Ravič 1967
10. Muller, S. 160 ff
11. Käding, 1897
12. Saukkonen, Haipus, 1979
13. Allén, 1971 und 1972
14. Rosengrén, 1972 und 1977
15. Michéa, S. 173 - 186, 314 - 332
16. Těšitelová, Otázky, S. 7
17. Muller, S. 164
18. Erk, 1972, S. 10
19. Das Sprach-Brockhaus
Deutsches Bildwörterbuch für jedermann,
Wiesbaden ⁷1962
20. Těšitelová, 1966
21. zitiert nach Hellmann, S. 26
22. Hellmann, S. 38
23. Materialy k častotnomy slovarju jazyka
Puškina, Moskva 1963
24. Rečnik na ezika na Christo Botev, sv. I Sofia
1960
25. Wenzel, Konkordanz zu den Dichtungen Georg
Trakls, Salzburg 1971; Klein, W., Zimmermann, H.
Index zu Georg Trakl Dichtungen, Frankfurt 1971
26. Těšitelová, 1974
27. Štejnfel'dt, o.J.
28. Jelínek, Bečka Těšitelová, S. 35 und 36
29. Zasorina, 1977
30. Těšitelová 1978, S. 56
31. Schröder, S. 219
32. zitiert nach Hellmann, S. 26
33. Těšitelová, 1977
34. Billmeier, S. 126 ff.
35. Bühler, K., Sprachtheorie, Jenä 1934
36. Mukařovský, S. 48
37. Morris, 1946

38. Kern, S. 6 und 8
39. Kučera, Francis, 1967
40. Wood, 1927
41. Alekseev, S. 62
42. Hoffmann, 1975, S. 35
43. Frumkina, 1964
44. Epstein, in: Josselson, 1953
45. Guiraud, 1960
46. Přihoda, S. 146 - 156
47. Frumkina

Kapitel II

1. Meister, S. 3
2. ebd., S. 61 - 63 und vgl. dazu auch Karlgren, 1967;
Sforza, Mailänder Herzogsgeschlecht, besaß 1450 -
1535 die Herrschaft über den großen Teil der
Lombardei
3. Kahn, S. 14
4. Käding, 1897
5. Bacmeister, 1870
6. Gamble, 1861
7. Knowles, 1904
8. Förstemann, S. 164
9. ebd., S. 165
10. Lord, S. 282
11. Mendenhall, 1887
12. Es gab keinen Mann des öffentlichen Lebens, den
der Publizist Junius nicht an der Wende der
60er und 70er Jahre des 18. Jahrhunderts ange-
griffen hatte. Die berühmte journalistische Lei-
stung dieses Junius besteht in 69 Briefen, die
der Herausgeber der Londoner Zeitung "Public
Advertiser", anonym erhalten, von 1769 bis 1772
veröffentlichte, ohne den wahren Namen seines
Mitarbeiters erfahren zu haben.
Vgl. dazu Kisch, E.E. (Hrsg.), Klassischer
Journalismus, Berlin 1923

13. Těšitelová, Nebeska, Králík, S. 119 - 147
14. Kädin, S. 5
15. Mistrík, 1969
16. Daniels, 1965
17. ebd., S. 274
18. Ortmann, 1975, 1976, 1978 und 1979
19. Ortmann, 1979, S. III
20. Rosengrén, 1972 und 1977
21. Hoffmann 1970, 1973, 1975 und 1976
22. Erk 1972, 1975 und 1976
23. Allén 1970 und 1971
24. Allén 1970, S. XIII
25. Frensch, Carter, Koenig, S. 290 - 324
26. Haarmann S. 156
27. Sedláček veröffentlichte später nur einen
Teil von seinen Auszählungen
28. erschienen in Prag, 1938
29. laut Mistrík 1969
30. Slovenský stenograf VIII, 7, 1954, S. 85 - 86
31. Stenografia I, Bratislava ³1955, S. 195 - 223
32. Mistrík 1969, S. 21
33. ebd., S. 112
34. Frumkina, vgl. Anm. 23, Kap. I
35. Josselson, 1953
36. Štejnfel'dt
37. Anemüller, 1974
38. ebd., S. 135
39. Radovich, 1974
40. Saukkonen, Haipus, 1979
41. Karvonen, Röman..., 1970
42. erschienen in 'Keelestatistika' 1, Tartu 1976
und 'Keelestatistika' 2, Tartu 1977
43. erschienen in 'Deutsche Kurzschrift' 1937

Kapitel III

1. Guiraud 1954
2. Mistrík 1969
3. ebd., S. 10
4. vgl. Zipf 1949; Zipfisches Gesetz, gilt nur für die Wörter mit mittlerer Frequenz und wird durch die Formel $r \cdot F = c$ definiert, wobei r (= Rang, Ordnungszahl einer bestimmten Worteinheit) multipliziert durch F (= Frequenz einer Worteinheit) ist konstant

Kapitel IV

1. Těšitelová 1973, S. 105 - 110
2. Saukkonen, Haipus..., S. 16
3. Es kommt auch vor, daß ein und dasselbe Wort als Postposition, als Präposition und sogar als Adverb gebraucht wird, z.B.
 - a) 'Tyttö käveli pihan poikki' (dt. 'das Mädchen ging durch den Hof') als Postposition
 - b) 'Tyttö juoksi poikki pihan' (dt. 'das Mädchen lief durch den Hof') als Präposition
 - c) 'Naru meni poikki' (dt. 'die Schnur brach ab') als Adverb
4. Mistrík 1973
5. Herdan untersuchte in 'Advanced Theory of Language as Choice and Chance' (Berlin-Heidelberg-New York 1966) die Distribution der Wortklassen am russischen Material
6. vgl. dazu Mistrík 1973

Kapitel V

1. Unter Wort wird hier eine lexikalische Einheit verstanden, die ein einzelnes Wort, eine Wortverbindung, eine grammatische Form oder eine von mehreren Bedeutungen eines Wortes sein kann.
2. Scheinpflugová, Olga, Český, Román, Praha 1969
3. ebd., S. 270 und 271
4. Joenpelto, Eeva, Kuin kekäle kädessä, Porvoo 1977
5. ebd., S. 166

Kapitel VI

1. Das Wort 'iltayö' tritt nicht im finnischen Frequenzwörterbuch auf. Es gehört zu den Wörtern, die in der Lexik des ganzen Korpus nur einmal vorkommen, insgesamt 58 % der Lexik. Alle diese Wörter wurden nicht ins Wörterbuch aufgenommen.
2. Mistrík 1969, S. 124
3. vgl. slovakisch, schwedisch, deutsch
4. Einwohnerzahl im Jahre 1970
5. Möglicherweise die höhere Frequenz durch die Nähe zu Prag zu erklären.
6. Einwohnerzahl im Jahre 1977

Schlußbemerkung

1. Saukkonen, Haipus..., 1979, S. 15
2. Mistrík, 1973, S. 26

LITERATURVERZEICHNIS

- ALEKSEEV, P.M.: Častotnye slovari i priemy
ich sostavlenija, Statistika
reči, Leningrad 1968
- ALLÉN, S.: Nusvensk frekvensordbok baserad
på tidningstext
1. Graford, homogrfkomponenter,
Stockholm 1970
----- Nusvensk frekvensordbok baserad på
tidningstext
2. Lemman, Stockholm 1971
----- Tiotusen i topp, Stockholm 1972
- ANEMOLLER, B.: Maschinelle Lexikographie. Ein
Beispiel aus dem Bereich der
Älteren slavischen Philologie.
Altbulgarische Übersetzung der
"Fragen und Antworten" des Pseudo-
Kaisarios. (Diss.) Saarbrücken,
1974
- BACMEISTER, A.: Stenotelegraphie (Wort- oder Buch-
stabenzählung?) in: Germanistische
Kleinigkeiten Stuttgart 1870,
S. 73 - 80
- BAKONYJ, H.: Die gebräuchlichsten Wörter der
deutschen Sprache, München 1934
- BAR-HILLEL, Y.: Die Zukunft der maschinellen Über-
setzung, oder: Warum Maschinen
das Übersetzen nicht erlernen.
in: Sprache im technischen Zeit-
alter, 1967, 23
----- The Present Status of Automatic
Translation of Language
in: Advances in Computers 1960,
1, S. 91 - 163 und 323 - 418
- BENSE, M.: Einführung in die Informations-
theoretische Ästhetik, Hamburg 1969
- BILLMEIER, G.: Über die Signifikanz von Auswahl-
texten, in: Forschungsberichte des
Instituts für deutsche Sprache 2,
Mannheim 1968, S. 126 - 171
- COHEN, M.: Sur la statistique linguistique,
Paris 1950
- DANIELS, K.: Deutsche Sprachstatistik.
in: Muttersprache 9, 1965 S. 273-279

- DIETRICH, R.,
KLEIN, W.: Computerlinguistik; Eine Einführung, Stuttgart 1974
- ERK, H.: Zur Lexik wissenschaftlicher Fachtexte Verben - Frequenz und Verwendungsweise, München 1972
- Zur Lexik wissenschaftlicher Fachtexte Substantive, München 1975
- Zur Lexik wissenschaftlicher Fachtexte Adverbien, München 1976
- FRENSCH, N.;
CARTER, C.;
KOENIG, W.: The Words and Sounds of Telephone Conversations, in: Bell Telephone Systems Technical Journal 1930, S. 290 - 324
- FÜRSTEMANN, E.: Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen, in: Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen. Hrsg. von Aufrecht, T. und Kuhn, A., Berlin 1852, Bd. 1, S. 163 - 179
- FRUMKINA, R.M.: Statističeskiye metody izučenija, Moskva 1964
- GAMBLE, W.: Two Lists of Selected Characters Containing all in the Bible and Twentyseven Other Books, Shanghai 1861
- GUIRAUD, P.: Problemes et méthodes de la statistique linguistique, Dordrecht/Holland 1960
- HAARMANN, H.: Balkanlinguistik 1, Tübingen 1978
- HARKIN, D.: The History of Word Counts in: Babel Vol. 3, No. 3, Bonn 1957
- HELLMANN, M.: Über Corpusgewinnung und Dokumentation im Mannheimer Institut für deutsche Sprache, in: Forschungsberichte des Instituts für deutsche Sprache 3, Mannheim 1969, S. 25 - 54

- HERDAN, G.: Quantitative Linguistics,
London 1964
- HOFFMANN, L.: Fachsprachen und Sprachstatistik,
(Hrsg.) Berlin 1975
- Fachwortschatz Physik, Häufig-
keitswörterbuch Russisch-Englisch-
Französisch, Leipzig 1970
- Fachwortschatz Chemie, Häufig-
keitswörterbuch Russisch-Englisch-
Französisch, Leipzig 1973
- Der häufigste Wortschatz der
Mathematik und EDV, Leipzig 1976
- Fachwortschatz Bauwesen,
Leipzig 1976
- JELÍNEK, J.; Frekvence slov, slovních druhů a
BEČKA, J.; tvarů v českém jazyce, Praha 1961
TEŠITELOVÁ, M.:
- JOSSELSOHN, H.H.: The Russian Word Count and Fre-
quency Analysis of Grammatical
Categories of Standard Literary
Russian, Detroit 1953
- Lexicography and the Computer,
in: To Honor Roman Jakobson.
Essays on the Occasion of His
Seventieth Birthday. The Hague-
Paris 1967, S. 1046 - 1057
- Automatic Translation of Language
Since 1960: A Linguist's View.
in: Advances in Computers 11,
1971, S. 1 - 58
- KAASIK, O.; Eesti keele ilukirjandusproosa
TULDAVA, J.; autorikõne sagedussõnastik
VILLUP, A.; (1. Teil), in: Keelestatistika 1,
XAREMAA, K.: Tartu 1976, S. 107 - 151
- Eesti keele ilukirjandusproosa
autorikõne sagedussõnastik
(2. Teil), in: Keelestatistika 2,
Tartu 1977, S. 141 - 169
- KÄDING, F.: Häufigkeitwörterbuch der deutschen
Sprache, Berlin-Steglitz 1897
- KAHN, D.: The Codebreakers. The Story of
Secret Writing New York 1968

- KARLGREN, H.: Statistical Methods in Phonetics, in: Manual of Phonetics. Edited by B. Malmberg Haag 1968, S. 129 - 152
- KARVONEN, J.; TAKALA, A.; RÖMAN, K.; YLINENTALO, O.: Opettajain sanastokirja, Jyväskylä 1970
- KERN, P.: Bemerkungen zum Problem der Textklassifikation, in: Forschungsberichte des Instituts für deutsche Sprache 3, Mannheim 1969, S. 3 - 33
- KNOWLES, R.: The London PointSystem of Reading for the Blind. London 1904
- KUČERA, H.; FRANCIS, W.: Computational Analysis of Present-Day-American English, Providence, Rhode Island 1967
- LORD, R.: Studies in the History of probability and statistics VIII. De Morgan and the statistical study of literary style, in: Biometrika 45, 1958, S. 282
- MENDENHALL, T.: The Characteristic Curves of Composition, in: Science 9, No. 214, 1887, S. 237 - 249
- A Mechanical Solution of a Literary Problem, in: The Popular Science Monthly 70, 1901, S. 97 - 105
- MEIER, H.: Deutsche Sprachstatistik, Hildesheim 1964
- MEISTER, A.: Die Anfänge der modernen diplomatischen Geheimschrift, Paderborn 1902
- MEL'ČUK, I.A.; RAVIČ, R.D.: Avtomatičeskij perevod, Moskva 1967
- MICHÉA, R.: Introduction pratique à une statistique du langage, in: Les Langues Modernes XLIII, 1949
- MISTRÍK, J.: Frekvencia slov v slovenčine, Bratislava 1969
- Exakte Typologie von Texten, München 1973

- MISTRÍK, J.: Frekvenčné zákonitosti v slovenčine
in: Studia Akademica Slovaca Bratislava 2, Bratislava 1973, S. 159 - 170
- Štylistika slovenského jazyka,
Bratislava 1970
- MORRIS, W.: Signs, Language and Behaviour,
New York 1946
- MUKAŘOVSKÝ, J.: Die poetische Benennung und die
ästhetische Funktion der Sprache,
in: Mukařovský, J., Kapitel aus
der Poetik, Frankfurt 1967,
S. 44 - 54
- MULLER, C.: Einführung in die Sprachstatistik
(dt.), Berlin 1972
- ORTMANN, W.: Hochfrequente deutsche Wortformen I,
München 1975
- Hochfrequente deutsche Wortformen II,
München 1976
- Hochfrequente deutsche Wortformen III,
München 1979
- Hochfrequente deutsche Wortformen IV,
München 1978
- PENDERGRAFT, E.: Translating Languages, in:
Borko, H. (Hrsg.) Automated Language
Processing, New York 1967, S. 291 - 323
- PIEPER, U.: Ober die statistischen Methoden für
die linguistischen Stilanalysen,
Tübingen 1979
- Differenzierung von Texten nach
numerischen Kriterien, in:
Folia Linguistica 1975, Bd. VII,
Den Hague, S. 61 - 113
- PŘÍHODA, V.: Rozvoj slovní zásoby, Gymnasium
36, 1948/49, S. 146 - 156
- RADOVICH, N.: Analisi Insiemistica Del Lessico
Slavo-Ecclesiastico Antico. Padua
1974
- ROSENGRÉN, I.: Ein Frequenzbuch der deutschen Zei-
tungssprache 1, Lund 1972
- Ein Frequenzbuch der deutschen Zei-
tungssprache 2, Lund 1977

- SAUKKONEN, P.;
HAIPUS, M.;
NIEMIKORPI, A.;
SULKALA, H.: Suomen kielen taajuussanasto
Porvoo 1979
- SEDLÁČEK, J.: Poměry slov a jejich částí v češtině,
získané rozbořem 18000 slov,
Těsnopisné rozhledy 2, 1924, S. 65 - 66
- SPRANG-HANSEN, H.: Mathematical Linguistics - A Trend
in Name or in Fact?
in: H.G.Lung (Hrsg.), Proceedings
of Ninth International Congress of
Linguistics. Cambridge, Mass.,
August 27 - 31, 1962 (= Janua
Linguarum, Series Maior 12) London -
The Hague, - Paris 1964
- ŠTEJNFEL'DT, E.: Häufigkeitswörterbuch der russischen
Sprache (dt.), Moskau, o.J.
Častotnyj slovar' sovremennogo
russkogo jazyka, Tallinn 1963
- SCHRÖDER, H.: Beziehung zwischen quantitativen
Stilmerkmalen und Intelligenz
und Charakter, in:
Muttersprache 72, 1962, S. 203 - 220
- TEŠITELOVÁ, M.: O morfologické homonymii v češtině
in: Rozpravy Československé Akademie
věd. Řada společenských věd, Roč.
76, Sešit 8, Praha 1966
- Zur Quantifikation der grammatischen
Kategorien, in: Linguistics 97,
Febr. 1/1973, The Hague, S. 105 - 110
- Otázky lexikální statistiky, Praha 1974
- Lingvistické příručky
Kvantitativní lingvistika, Praha 1977
- Frekvence slov a tvaru ve spise
"Život a dílo skladatele Foltýna",
od Karla Čapka, Naše řeč 32, 1948,
S. 126 - 130
- On some Questions of Spoken
Scientific Discourses of Men and Women
(From Point of View of Quantitative
Analysis of their Vocabulary)
in: PSML 6, Prague 1978, S. 47 - 58
- TEŠITELOVÁ, M.;
NEBEŠKA, I.;
KRÁLÍK, J.: On the Quantitative Characteristics
of the Czech Texts of Disputed
Authorship - RKZ, in: PSML 5, Prague
1976, S. 119 - 147

- WOOD, B.: A Comparative Study of the
Vocabularies of Sixteen French
Textbooks, Reprinted from the
Modern Languages Journal 11,
1927/5
- ZASORINA, L.N.: Častotnyj slovar' russkogo jazyka,
(Hrsg.) Moskva 1977
- ZIPF, G.: Human Behavior and the Principle
of Least Effort, Cambridge 1949
- PSML = Prague Studies in Mathematical
Linguistics. Praha

LEBENS LAUF

Am 28.12.1946 wurde ich in Sippola/Finnland geboren. Nach vierjährigem Besuch der Volksschule und achtjährigem Besuch der Oberschule in Myllykoski legte ich im Mai 1965 die Reifeprüfung ab.

An den Universitäten Helsinki und Freiburg i. Br. studierte ich Slavische Philologie und Neuere deutsche Literaturgeschichte. Im Juli 1975 legte ich das Magister-Artium-Examen ab.

In der Zeit von November 1977 bis September 1978 arbeitete ich im Erzbischöflichen Archiv und anschließend bis Mai 1980 im Sprachlabor der Universität, beide in Freiburg i. Br.

Meine Lehrer waren die Professoren und Dozenten Kiparsky, Lettenbauer, Měšťan, Vahros, Weiher, Dyck, Schwarz, Steger, Kochendörfer, Lohr.

Herrn Professor Dr. A. Měšťan bin ich für Rat und wohlwollende Förderung besonderen Dank schuldig.

